АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ ХАРАКТЕРИСТИК ЛИЧНОСТИ АВТОРА НА ОСНОВЕ АНАЛИЗА СООБЩЕНИЙ В СОЦИАЛЬНЫХ СЕТЯХ

Зефирова Татьяна Викторовна (tatyana.zefirova@gmail.com)
Лукашевич Наталья Валентиновна (louk_nat@mail.ru)

МГУ им. М. В. Ломоносова, Москва

Annotation
Automatic processing methods allow to extract personal linguistic and extra-linguistics traits that are valuable in terms of distinguishing certain classes of Internet users. The goal of our research was to define which user characteristics may be important for such investigations in regard to a binary scale of introversion/extraversion which has been earlier proved to be one of the most stable ones in popular personality models. The research was performed on public materials of Twitter users self-identifying themselves with either ones of binary psychological prototypes (introvert/extrovert). Obtained results on the lexical layer of the texts contribute to creating a lexical portrait of a typical introversive/extroversive social nets user.

Keywords: automatic feature extraction, statistical analysis, natural language processing, machine learning, psycholinguistics.

Автоматическое определение характеристик личности автора на основе созданных им текстов открывает широкий спектр прикладных возможностей. Автоматический анализ публичных сообщений в блогах и социальных сетях может помочь выявить психически нестабильных людей, социально опасных или нуждающихся в помощи.

Исследования такого рода могут ставить перед собой задачу отнести автора сообщений к одному из заранее определенных типов личностей. Определение таких типов является "психологической" стороной задачи анализа текстов. Наиболее используемыми являются системы FFM (Five Factor Model, The Big Five) и MBTI (Myers-Brigg Type Indicator, шкала Майерс-Бриггс). Самые стабильные результаты показывает общая для этих двух систем шкала интроверсии/экстраверсии, поэтому именно эта шкала была выбрана для использования в данном исследовании.

Сообщения для анализа были собраны через сеть мини-блогов Twitter. Этот выбор обусловлен несколькими факторами. Во-первых, сообщения в социальных сетях создаются авторами в "экологических" условиях, в отличие от текстов, созданных при работе со специалистами-психологами при выполнении различных тестов. Во-вторых, сообщения доступны практически в неограниченных количествах, а средства для разработчиков Twitter позволяют легко осуществить автоматический сбор данных для наполнения баз.

Для того чтобы создать базы пользователей Twitter, являющихся интровертами или экстравертами, был проведен автоматический поиск авторов, самоидентифицирующих себя соответствующим образом, то есть называющими себя так в собственных сообщениях. После получения базы данных автоматическим образом все же требовалось провести ручную проверку на адекватность ее заполнения. Больша́я часть полученных сообщений удалялась. Сомнительными считались следующие категории:

1) Ретвиты
2) Сообщения, не являющиеся самоидентифицированием
3) Неопределенные твиты (содержащие упоминание более чем одного типа и другие)
4) Твиты на иностранных языках (украинский, македонский…):

Кроме того, в базу заносились с метаданные: общее количество сообщений, количество подписчиков, количество подписок. К именам пользователей и текстам сообщений было применено условие уникальности, таким образом отсекалась часть ретвитов и было исключено попадание в базу нескольких входов от одного и того же пользователя.

Следующим этапом стал сбор последних 100 сообщений от каждого из пользователей, оставшихся в базе. По результатам этого сбора также пришлось удалить часть ранее сохраненных входов в базу данных, поскольку среди аккаунтов оказывались не персональные (новостные, фальшивые и т.д.). Кроме того, часть аккаунтов к определенному моменту работы оказывалась в ограниченном пользователями доступе, поэтому дальнейшая работа с ними была невозможна. На этом этапе также удалялись ретвиты. Из самих текстов сообщений были удалены упоминания других пользователей, ссылки на сторонние ресурсы, ссылки на изображения, часть хэштегов (за исключением находящихся внутри текста или напрямую соответствующих основному тексту сообщения), географические отметки, нетекстовые смайлы (не состоящие из стандартных знаков препинания и символов алфавита). В данном случае сообщения на других языках, кроме русского, сохранялись, впрочем, такие сообщения оказались статистически незначимыми.

Для анализа были использованы средства стандартного пакета средств машинного обучения scikit-learn для создания программ автоматической обработки полученного

корпуса на языке Python. Для извлечения важных текстовых характеристик использовалась статистическая мера TF*IDF. Анализ проводился на основе n-грамов с $n \in 1, 2$. Хотя при исследованиях на материале английского языка исключение стоп-слов было сочтено ухудшающим результаты, на материале русского языка оно оказалось необходимым.

По результатам анализа были получены следующие списки 20 слов с наибольшими весовыми коэффициентами для групп интровертов и экстравертов.

extra			intro		
occ	term	weight	occ	term	weight
7172	это	0.361967	5922	это	0.198766
2521	кофе	0 238898	701	вообще	0.070178
4840	просто	0.166505	5562	хаха	0.066653
6195	тебе	0.152026	4493	сегодня	0.064347
5392	сегодня	0.152026	4068	просто	0.057083
2832	люблю	0.115830	814	всё	0.049541
4586	почему	0.115830	5669	хочу	0.049196
3964	очень	0.108590	1393	ещё	0.046567
5803	спасибо	0.108590	1465	жизни	0.046240
1406	день	0.094112	1123	делать	0.042369
6837	хочу	0.094112	5197	тихо	0.040942
465	блин	0.094112	1480	жить	0.038179
581	буду	0.086872	2523	мир	0.036396
938	время	0.086872	4851	спать	0.034569
6096	такая	0.079633	3331	очень	0.033799
2861	людей	0.072393	3839	почему	0.033799
1488	Дня	0.072393	4366	рядом	0.033767
1708	ещё	0.065154	2672	наверное	0.033422
899	вообще	0.065154	1710	знаю	0.033422
3327	настроение	0.065154	4837	спасибо	0.031112

Поскольку часть слов в обоих списка совпадает, было принято решение о том, что их можно считать своеобразными стоп-словами именно для материала социальных сетей (такие слова как «сегодня», «хочу», «спасибо» достаточно показательны в отношении содержания типичного сообщения в мини-блоге) и исключить из списка.

В результирующем списке были выделены слова, попадающие в такие категории, определяющие речь интровертов и экстравертов (согласно F. Mairesse, M. A. Walker, M. R. Mehl and R. K. Moore (2007) [9]), как social words (жирным) и positive words (курсивом):

extra			intro		
occ	term	weight	occ	term	weight
2521	кофе	0 238898	5562	***хаха***	0.066653
6195	**тебе**	0.152026	814	всё	0.049541
2832	***люблю***	0.115830	1465	жизни	0.046240
1406	день	0.094112	1123	делать	0.042369
465	**блин**	0.094112	5197	тихо	0.040942
581	буду	0.086872	1480	жить	0.038179
938	время	0.086872	2523	мир	0.036396
6096	такая	0.079633	4851	спать	0.034569
2861	**людей**	0.072393	4366	рядом	0.033767
1488	Дня	0.072393	2672	наверное	0.033422
3327	настроение	0.065154	1710	знаю	0.033422

В соответствии с традиционными представлениями о такой черте личности как экстраверсия, слова, которые можно отнести в категорию social words, в достаточном количестве представлены в списке статистически значимых слов для базы текстов экстравертов. К «социальным словам» нами был также отнесем такой элемент как «блин», который можно отнести к маркерам устного разговорного дискурса.

Отдельно отметим единственное «социальное» слово в списке статически значимых для текстов интровертов слов – «хаха». С одной стороны, эта форма, которую можно отнести к социальным и положительно эмоционально окрашенным, что вступает в противоречие с остальными наблюдениями относительно полученных списков. С другой стороны, при ручной проверке текстов пользователей было отмечено большое количество разнообразных форм написания аналогичных выражений, поэтому точно оценить встречаемость междометия и, соответственно, его статистическую значимость в текстах интровертов и экстравертов сложно.

Интересно отметить, что единственным значимым словом, получившим весовой коэффициент достаточный даже для попадания с список без подключения стоп-слов, оказалось слово «кофе» в списке для текстов экстравертов.

Если же рассмотреть список полученных слов, статистически значимых для текстов интровертов, можно обратить внимание, что в него попадают такие слова как «жизни», «тихо», «жить», «мир», «спать», которые в действительности имеют менее «социальные» коннотации.

Для более подробного анализа категорий слов полученные слова были распределены по категориям, предложенным в Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Inquiry and Word Count: LIWC 2001. Lawrence Erlbaum, Mahwah, NJ. (полный список категорий в Pennebaker at al. 2007 [43]). При этом составляющие списков значимых слов для текстов интровертов и экстравертов находятся практически в дополнительном распределении (то есть категория, заполненная для одной группы, скорее всего не заполнена для второй). Отметим, что в пересекающиеся категории попадают элементы «хаха» и «кофе», появление которых в списке уже было прокомментировано отдельно.

На основании полученного распределения можно сделать предположение о том, что тексты интровертов и экстравертов, публикуемые ими в социальных сетях, действительно различаются с точки зрения семантики и лексики. Более подробное исследование (на более обширном и/или проконтролированном материале) может позволить создать «языковой портрет» русскоязычного пользователя социальных сетей, определяющего себя как интроверта или экстраверта. Отметим также, что получившийся на данный момент «портрет» вполне соответствует наивным представлениям о типичном экстраверте или интроверте. Так, в текстах интровертах существенными оказываются слова обозначающие когнитивные процессы («знаю»), общие глаголы и глаголы, обозначающие физические состояния («спать», «жить»), в то время как среди значимых слов для экстравертов оказываются слова с положительными эмоциональными коннотациями («люблю»), а также, что интересно, сразу несколько слов, попадающих в категорию описания времени – «время», «день» (а также форма «дня»).

Для классификации и обучения модели были выбраны для сравнения два алгоритма классификации: наивный Байесовский классификатор и метод опорных векторов (SVM, Support Vector Machines).

Точность предсказания наивного Байесовского классификатора составила 0.62. Точность предсказания при использовании метода опорных векторов оказалась выше – 0.65. Заметим, что при увеличении объема используемого корпуса точность предсказания предположительно будет повышена.

Проведенный дополнительно анализ Grid Search (модуль GridSeacrhCV в scikit-learn) показал, что при выборе коэффициента экспоненциального сглаживания alpha = 0.01 и ограничении размера n-граммов до n=1, а также увеличения числа итераций метода стохастического градиента со стандартных 5 до 8 полнота работы классификатора SVM достигает уровня 80%.

Сравнение работы классификаторов:

	precision	recall	f1-score
NBC total	0.62	0.52	0.46
SVM total	0.65	0.59	0.54
SVM-tuned	0.54	0.80	0.64

Отдельно были проанализированы метаданные. Среди метаданных мы рассматривали общее количество сообщений, количество подписок (друзей) и количество подписчиков. Судя по полученным результатам, несмотря на исходное предположение о том, что интроверты предпочитают общаться в Интернете, а, следовательно, пишут больше сообщений в социальных сетях, среднее количество сообщений у самоопределившихся интровертов заметно меньше. Среднее количество статусов на странице в Twitter у интроверта (mean = 17990.189189) почти в два раза ниже, чем на странице экстраверта (mean = 35404.666667). Отметим также, что значительно ниже у интровертов оказываются и границы области количества сообщений. Число статусов у пользователя-экстраверта лежит в области от 31 до 364527 сообщений, в то время как аналогичная область у пользователя-интроверта ограничена 20 и 190689 сообщениями соответственно.

Итогами проведенной работы стали:

1. Создание баз текстов и метаданных пользователей социальной сети мини-блогов Twitter, самоидентифицирующих себя как принадлежащих к одному из типов по шкале интроверсии/экстраверсии;
2. Анализ текстовых данных полученных баз, включая:
    - ручной анализ лексических составляющих;
    - автоматическое определение весов составляющих в модели «мешок слов»;
    - анализ полученных списков наиболее статистически значимых слов для текстов обоих баз;
3. Создание алгоритма классификации документов и оценка его работы;
4. Анализ полученных метаданных.

Проделанная работа может служить отправной точкой для новых исследований, в частности, в области сравнения результатов с контрольными группами интровертов и экстравертов, увеличения баз данных для уточнения результатов классификатора, а также изучения значимости других метаданных или более подробного анализа полученных текстов на синтаксическом и семантическом уровне.

Список основной использованной литературы:

[1] Bogdanova, D., Rosso, P., & Solorio, T. (2014). Exploring high-level features for detecting cyberpedophilia. Computer Speech & Language

[2] Dewaele, J.-M., & Furnham, A. (1999). Extraversion: the unloved variable in applied linguistic research. Language Learning, 49 (3), 509–544.

[3] Furnham, A., & Mitchell, J. (1991). Personality, needs, social skills and academic achievement: A longitudinal study. Personality and Individual Differences, 12, 1067–1073

[4] Gill, A. (2003). Personality and Language: The Projection and Perception of Personality in Computer-Mediated Communication. Ph.D. thesis, University of Edinburgh.

[5] Gill, A. J., & Oberlander, J. (2002). Taking care of the linguistic features of extraversion. In Proceedings of the 24th Annual Conference of the Cognitive Science Society, pp. 363–368. Goldberg, L. R. (1990). An alternative “description of personality”: The Big-Five factor structure. Journal of Personality and Social Psychology, 59, 1216–1229.

[6] Hirsh, J.B., & Peterson, J.B. (2009). Personality and language use in self-narratives. Journal of Personality in Research, 43, 524-527.

[7] Inches, G., Crestani, F. (2012). Overview of the international sexual predator identification competition at pan-2012. In CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers, eds. P. Forner, J. Karlgren, C. Womser-Hacker (Rome, Italy).

[8] John. O. P. (1990). The "Big Five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. Pervin (Ed.), Handbook of personality theory and research (pp. 66-100). New York: Guilford

[9] Mairesse F., M. A. Walker, M. R. Mehl and R. K. Moore (2007) "Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text", Volume 30, pages 457-500

[10] Mairesse, F., & Walker, M. (2006). Automatic recognition of personality in conversation. In Proceedings of HLT-NAACL.

[11] McCrae, R. R. & John, O.P. (1992). An introduction to the five-factor model and its applications. Journal of Personality, 60(2), 175-215.

[12] Mehl, M. R. (2006). Quantitative text analysis. Handbook of Multimethod Measurement in Psychology, 141-156.

[13] Neuman Y., Personality Research for NLP, A tutorial presented at the 2015 Conference on Empirical Methods on Natural Language Processing, September, Lisbon, Portugal.

[14] Pennebaker James W., Chung Cindy K., Ireland Molly, Gonzales Amy, and Roger J. Booth. The Development and Psychometric. Properties of LIWC2007.

[15] Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Inquiry and Word Count: LIWC 2001. Lawrence Erlbaum, Mahwah, NJ.

[16] Plank, Hovy, Personality traits on Twitter. Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015), pages 92–98,

[17] Scherer, K. R. (1979). Personality markers in speech. In Scherer, K. R., & Giles, H. (Eds.), Social markers in speech, pp. 147–209. Cambridge University Press.