Department of Information Engineering and Computer Science

Bachelor's degree in
Computer Science

Final Dissertation

TITLE TODO

Supervisor
Alberto Montresor
Co-Supervisor
Daniele Miorandi

Student
Stefano Perenzoni

Academic year 2019/2020

Contents

Abstract

Abstract

1 Introduction

During my internship at U-Hopper, I had the opportunity to develop this Thesis as a result of my experience inside the company. *U-Hopper is a research-intensive deep-tech SME, headquartered in Trento, providing big data-enabled solutions and technologies for the government, retail and manufacturing sectors. U-Hopper has received numerous awards for its innovative solutions, including, among the others, the Lamarck prize (2013), a EC Seal of Excellence (2015), the Innov@Retail prize (2016) and a nomination for the 2017 EC Innovation Radar Awards.*. The company is active in many different domains such as retail and tourism and offers a variety of competences including chatbots, analytics, and machine learning. Thanks to Tapoi[1], an innovative data intelligence solution, U-Hopper is also into the sector of user profiling. It allows businesses to deliver personalized experiences to their customers through the mining and analysis of their activities on social networks. Thus, the extraction of behavioural insights can be a valuable aspect since being aware of how an individual comes to a decision helps to provide each customer with the right tailored content.

1.1 Motivation and business requirements

Dissatisfied customers represent a dangerous threat for companies and their brands. Thus, it is fundamental for a business to track audience satisfaction and do whatever it can to fulfil their want. Dissatisfaction can impact a company in two different ways. First, those who are not completely satisfied would behave passively towards the business, reducing the number of purchases, and therefore stop being consumers of its products and services. Moreover, those who are more active and extroverted could interact with others and convey their disappointment. Overall, a large number of unhappy customers will entail a significant loss of customers.

This problem is of particular interest to those typologies of companies that follow a *business-to-customer (B2C)* sales process, with a wide customer base and which interactions with their audience are characterized by online relationships. This relation can be purely telematic, as in the case of e-commerce, or it can support a physical one where the material interaction is unavoidable, as in the case of banking and insurance sectors.

For this kind of businesses, customers' satisfaction is not trivial to accomplish since each one of them has different needs and requests and standard methodologies do not adapt well for everyone. Thus, over the past few years, personalization of customer experience has become vital in order to inspire an honest and natural emotional response. It is then important to be able to access information which allows marketeers to offer fully tailored contents, through a specific mean of communication and with personalized messages to meet each individual's requirements.

While, thanks to Customer Relationship Management (CRM), data related to the direct interaction between customer and company has already been deeply explored, social media networks gave access to more personal information allowing a deeper understanding of the person. The system discussed in this thesis proposes a solution that goes further than the diffused purchase history-based personalization. It aims to provide companies with the ability to extract readable and valuable insights about singular individuals from their activities online. The final goal is to make available actionable insights about users' behaviour, demographics and attitudes. In particular, this dissertation focusses on the extraction of personality traits

[1] `www.tapoi.me`

to obtain a detailed description of a person's behaviour and reaction to a number of observed solicitations

1.2 Customer insights

General introduction to customer insights, different types, benefit for businesses
At the end, a short focus on psychometrics

1.3 Extraction of personality models

According to neuroscientists Adelstein et al., personality describes human behavioural responses to wide classes of external stimuli [2]. It works as an adaptive system for taking in, organizing information and driving the response to inner and outside demands [7]. The parameters of the adaptive system represent the variation of the same from person to person and, therefore, characterize uniquely every individual. These parameters are also referred to as personality traits in several different personality models studied over the years. Each model includes its range of traits which combinations describe several personality types. Researchers have shown clear connections between general personality traits and many types of behaviour.

Some fundamental traits describe the type of relationship a person has with the outside world and the way he or she communicates [27]. Thus, to facilitate communication, recently, businesses are using personality models to gain a better understanding of what drives the interests of a person. This approach is showing clear benefits in many different applications. In the field of Human-Computer Interaction, users prefer interfaces designed to represent personalities that most closely matched their own [33]. Some studies have also suggested connections between customer personality and marketing. Through techniques more focused on the target audience, it is possible to profile individuals, and tailor advertisement automatically displayed based on their personality [5]. Therefore, the ability to identify people's personality or, even better, details of their personality traits through well-defined models is a significant competitive advantage since we would have a precise representation of the customer's reasoning process.

1.3.1 Big Five personal traits

While several models exist, the *Big Five*, also known as the *five-factor model* and the *OCEAN model* is one of the most well-researched and widely accepted taxonomies among scientists [30, 29]. It formalizes personality along 5 domains, namely Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism. Each one of these traits is continuous and usually ranges on a scale from 1 to 5. High openness marks imagination, creativity, and curiosity in learning and exploring new things. Conscientiousness represents self-discipline and attention to details. Extroversion measures preferences for interacting with other people. Agreeableness reflects the extent to which a person is generous, trustworthy and always willing to help others. Finally, a high score on neuroticism indicates a tendency to get stuck in negative emotions. At the two extremes of each trait, two separate aspects reflect a particular behaviour. For example, conscientiousness is bounded by carelessly at the lowest end and by organization and efficiency at the greatest one.

Since its first definition, this model rapidly became one of the standards in the psychological community, largely accepted by the most share of scientists since it allows to describe accurately the traits of a singular. However, concerning the exploitation of personality information in the work and marketing environments, it received some critics about the extraction of actionable insights[21, 35]. Indeed, since each trait is represented by a real number between 2 extremes, it has been argued to be hardly readable and therefore less valuable for fields such as marketing and business. Thus, structures based on clearer distinctions are often preferred.

1.3.2 Myers-Briggs Type Indicator

The Myers-Briggs model, also called *Myers-Briggs Type Indicator, or MBTI*, is the most common alternative to the Big-Five model. Contrarily to the former, there are discussion about the MBTI and its limitations in reflecting the whole personality system. Boyle and Barbuto are two of the scientists that presented a number of psychometric limitations pertaining to the validity and reliability of this model [8, 6]. However, many of their arguments have been proved wrong by Furnham who demonstrated several correlations between the dimensions defined by Myers and the big five factors [17].

The MBTI is a categorical model, based on the conceptual theory of Jung and developed by Katharine Briggs and Isabel Myers who used four different dichotomies to evaluate the personality of people [23]. A first one differentiates a person's attitude in either extraversion (E) or introversion (I).These two preferences describe if one focusses on external stimuli, such as action and interaction with other people or internal ones like self-reflection. Two perceiving functions, sensation (S) and intuition (N) describe the process of gathering new information. On the one hand, people who trust tangible and concrete facts; on the other hand, those who tend to find patterns and meaning also regarding future possibilities. The third cognitive function is that of decision-making which can be thinking (T) or feeling (F). While thinkers make reasonable and consistent choices and reflect over consequences applying a rigid set of rules, feelers tend to emphasize with the situation considering the needs of people involved Finally, there is the lifestyle preference function dichotomy, judging (J) or perceiving (P). Judging types like the outside world to be structured; according to Myers, they prefer to "have maters settled". On the contrary, perceiving personalities like it flexible and spontaneous and tend to "keep decisions open" [32]. There are 16 different types of personality given by the combination of these 4 cognitive functions identified by 4-characters codes such as "INFJ" or "ENFP".

Using a categorical model, the extraction of personality from social media activities is a *machine learning* problem, precisely, it consists of numerous classification tasks, one for each of the four variables. Machine learning is one of the most talked-about fields of computer science and many sources give their own definition. Basically, ML deals with allowing a computer system to "learn with data, without being explicitly programmed" [42]. It has been applied in many contexts, such as decision making, optimization problems, forecasts, and predictions. Nowadays, we face ourselves with machine learning in everyday life: home assistants, security surveillance, music and shopping suggestions, customer services are strongly powered by artificial intelligence. These services rely on data to learn how to work as good as possible: they are trained with samples of data similar to what they expect to receive by their users: the more accurate, exhaustive and in large quantities they are, the better the system learns. Therefore, data have a very central role in machine learning problems.

A classification task has the goal of assigning a belonging class to a given object. The input is composed by a tuple of *features* that characterize the object, usually made by numbers, and the output is a categorical variable, such as a "yes/no" label. In other words, it can be seen as a mathematical function, that maps a vector $\boldsymbol{x} \in \mathbb{R}^n$ to an answer $y \in C$

$$f \colon \mathbb{R}^n \to C$$
$$f \colon \boldsymbol{x} \mapsto y$$

where C is a set of possible categories. For example, in one of the four classifiers for this problem, $\boldsymbol{x}$ represents a user and her activities on the social media, and $C = \{\texttt{Introvert}, \texttt{Extrovert}\}$

1.4 Research objectives

Extraction of behavioural insights from social media has recently attracted the attention of both researchers and businesses. Even though the latter has released a couple of solutions, these fit better for personal and psychological use rather than a commercial one. The main objective of this thesis is to design and develop a solution that can be used by a company to personalize

customer experience with respect to individual abstract preferences. Therefore, the question it answers is: *is it possible to understand costumers behaviour from their online profiles and activities?*

The designed system should be able to work with numerous social media platforms to have a wide variety of data sources. Finally, the principal aspect that it must always satisfy is the *ability to use the result*. Indeed, extracted insights need to be actually actionable, directly by the marketing department or in conjunction with further analysis, to represent a competitive advantage.

1.5 Outline

Chapter 2 describes the state of the art. Chapter 3 introduces the design of the solution. It focuses on used components and algorithms, their logic and their interfaces. Chapter 4 shows how the mentioned components are implemented and integrated. It follows the implementation of the algorithm and the evaluation of a general prototype of the proposed system. Chapter 5 concludes the thesis with some observations and future work proposals.

2 State of the Art

This chapter presents the current state of the art regarding insights extraction on social media. Many aspects of online users have been explored in order to profile customers. "Then, there is a focus on what has been done in terms of providing actionable personality insights."

Some studies aimed to identify clear demographic characteristics based on both the analysis of a user's activities and her network inside the social media. Twitter is commonly used for the extraction of gender [31], age or age groups [11]. Also, a person's family status is inferred through the detection of life events such as the birth of a child and a marriage [13].

The literature also presents many examples of latent attributes extraction. Some of the most remarkable research has been carried out by the *World Well-Being Project* [2]; a research center which used social media to measure attitudes and personal characteristics such as optimism and pessimism [41], temporal orientation [43]. Many different social networks have been explored as well as many aspects that are not limited only to text but also include images and social interactions. Finally, it is a common practice inferring behaviour through a variety of personality models.

However, what has been done is almost completely focused only on the feasibility of extracting attitudes' insights from online activities rather than a commercial use of the obtained information to generate a marketing advantage. So, the literature presents only a few systems which satisfy the right requirements for an application in the real world, such as those imposed by the GDPR TODO CITE.

2.1 Customers Profiling

A precise and detailed description of social media users requires the analysis of many aspects of social media. Indeed, understanding the users means being able to quantify and qualify how they present themselves [44].

Many of the systems proposed for social media analysis use as fundamental component features that describe interactions of users, such as the number of followers, mentions, likes, and comments. This type of analyses has been largely explored since studies about user influence and social engagement. First, raw measures publicly available on social media were used to calculate metrics to represent effectively the user's influence [34]. Further research proved that simply observing ground numbers of a profile can lead to a misunderstanding. Cha stated that the indegree alone (number of followers) reveals little and suggested to consider shares and mentions from other users [9]. D. Romero et al. observed influence analysing the propagation of web links over time using both the structural properties of the network as well as the diffusion behaviour among users [40]. They also regarded the *passivity of a user*, a measure of how difficult it is for other users to influence him, and used it to weigh the tweets propagation network. Many different networks can be explored on social media in order to identify influence, communities, and trend topics applying the myriad of network concepts and analyses such as degree centrality and modularity [10]. The nature of these graphs can change regarding the platform's characteristics and the aspect we are looking at. Li complained about undirected networks, such as the Facebook friends graph and proposed a method based on the *Share/Reply/Mention* directed network to

[2] https://wwbp.org/

capture user influence [26]. These observation are usually used to profile a person's social environment and to assess his or her role inside it.

A second fundamental point carried out by literature on social media is the analysis of the context the user is talking about. Obviously, being aware of what topics drive someone's interactions is essential to profile his or her interest. Moreover, they can be used to reduce other types of analyses to a specific field of interest. For example, focussing on users' influence in sports discussions. To understand context, it is necessary to observe the content of the messages which is usually composed by text and images or videos. Firstly, keywords in the activities were used to identify topics [9]. This methodology shows some clear issues, especially when used for social media when messages tend to be extremely abbreviated through acronyms and slang words. Other approaches, feasible in a limited number of platforms, proposed to use most used hashtags to obtain linguistic content starting from the activities [36]. Finally, a more general technique is using the tree of Wikipedia categories to characterize the user's interests. This method fits well with both text and multimedia content thanks to a number of services that apply semantic analysis techniques to extract relevant entities [46].

2.2 Personality insights

"Psychometric profiling is the process by which your actions are used to infer your personality."

The analyses listed before are essential when it comes to extraction of behavioural insights. We can understand how a person behave and react to external stimuli Furthermore, studies have shown a strong correlation between discussed topic and personality aspects.

Far vedere diversi tipi di analisi fatti per la personalitÃ Poi dire che molti si basano sul big5 e far vedere quelli sull'MBTI

2.3 Commercial applications

3 Design and methodology

3.1 Components logic

3.2 Algorithms

4 Implementation

4.1 Components interactions

4.2 Algorithms implementation

5 Evaluation

5.1 Evaluation metrics

5.2 Performance evaluation of the system

6 Conclusions

6.1 Limitations

6.2 Future work

References

[1] Muhammad M Abdul-Mageed et al. "Recognizing pathogenic empathy in social media". In: *Eleventh International AAAI Conference on Web and Social Media*. 2017.

[2] Jonathan S Adelstein et al. "Personality is reflected in the brain's instrinsic functional architecture". In: *PloS one* (2011).

[3] Yair Amichai-Hamburger and Gideon Vinitzky. "Social network use and personality". In: *Computers in human behavior* 26.6 (2010), pp. 1289–1295.

[4] Danny Azucar, Davide Marengo, and Michele Settanni. "Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis". In: *Personality and individual differences* 124 (2018), pp. 150–159.

[5] Yoram Bachrach et al. "Personality and patterns of Facebook usage". In: *Proceedings of the 4th annual ACM web science conference*. 2012, pp. 24–32.

[6] John E Barbuto Jr. "A critique of the Myers-Briggs Type Indicator and its operationalization of Carl Jung's psychological types". In: *Psychological Reports* 80.2 (1997), pp. 611–625.

[7] Jack Block. *Personality as an affect-processing system: Toward an integrative theory*. Psychology Press, 2002.

[8] Gregory J Boyle. "Myers-Briggs type indicator (MBTI): some psychometric limitations". In: *Australian Psychologist* 30.1 (1995), pp. 71–74.

[9] Meeyoung Cha et al. "Measuring user influence in twitter: The million follower fallacy". In: *fourth international AAAI conference on weblogs and social media*. 2010.

[10] Bongsug Kevin Chae. "Insights from hashtag# supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research". In: *International Journal of Production Economics* 165 (2015), pp. 247–259.

[11] Aron Culotta, Nirmal Ravi Kumar, and Jennifer Cutler. "Predicting the Demographics of Twitter Users from Website Traffic Data." In: *AAAI*. Vol. 15. Austin, TX. 2015, pp. 72–8.

[12] Colin G DeYoung. "Toward a theory of the Big Five". In: *Psychological Inquiry* 21.1 (2010), pp. 26–33.

[13] Thomas Dickinson et al. "Identifying prominent life events on twitter". In: *Proceedings of the 8th International Conference on Knowledge Capture*. 2015, pp. 1–8.

[14] Golnoosh Farnadi et al. "A multivariate regression approach to personality impression recognition of vloggers". In: *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*. 2014, pp. 1–6.

[15] Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. "Predicting personality traits with instagram pictures". In: *Proceedings of the 3rd Workshop on Emotions and Personality in Personalized Systems 2015*. 2015, pp. 7–10.

[16] John W Foreman. *Data smart: Using data science to transform information into insight*. John Wiley & Sons, 2013.

[17] Adrian Furnham. "The big five versus the big four: the relationship between the Myers-Briggs Type Indicator (MBTI) and NEO-PI five factor model of personality". In: *Personality and Individual Differences* 21.2 (1996), pp. 303–307.

[18] Jennifer Golbeck, Cristina Robles, and Karen Turner. "Predicting personality with social media". In: *CHI'11 extended abstracts on human factors in computing systems*. 2011, pp. 253–262.

[19] Jennifer Golbeck et al. "Predicting personality from twitter". In: *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE. 2011, pp. 149–156.

[20] Sharath Chandra Guntuku et al. "Studying personality through the content of posted and liked images on Twitter". In: *Proceedings of the 2017 ACM on web science conference*. 2017, pp. 223–227.

[21] Leaetta M Hough and Adrian Furnham. "Use of personality variables in work settings". In: *Handbook of psychology* (2003), pp. 131–169.

[22] Kokil Jaidka, Sharath Chandra Guntuku, and Lyle H Ungar. "Facebook versus Twitter: Differences in Self-Disclosure and Trait Prediction". In: *Twelfth International AAAI Conference on Web and Social Media*. 2018.

[23] Carl G Jung. "Personality types". In: *The portable Jung* (1971), pp. 178–272.

[24] Margaret L Kern et al. "Gaining insights from social media language: Methodologies and challenges." In: *Psychological methods* 21.4 (2016), p. 507.

[25] Michal Kosinski, David Stillwell, and Thore Graepel. "Private traits and attributes are predictable from digital records of human behavior". In: *Proceedings of the national academy of sciences* 110.15 (2013), pp. 5802–5805.

[26] Jingxuan Li et al. "Social network user influence sense-making and dynamics prediction". In: *Expert Systems with Applications* 41.11 (2014), pp. 5115–5124.

[27] Ana Carolina ES Lima and Leandro N de Castro. "Predicting temperament from Twitter data". In: *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*. IEEE. 2016, pp. 599–604.

[28] Fei Liu, Julien Perez, and Scott Nowson. "A language-independent and compositional model for personality trait recognition from short texts". In: *arXiv preprint arXiv:1610.04345* (2016).

[29] Robert R McCrae and Paul T Costa. "Validation of the five-factor model of personality across instruments and observers." In: *Journal of personality and social psychology* 52.1 (1987), p. 81.

[30] Robert R McCrae and Oliver P John. "An introduction to the five-factor model and its applications". In: *Journal of personality* 60.2 (1992), pp. 175–215.

[31] Zachary Miller, Brian Dickinson, and Wei Hu. "Gender prediction on twitter using stream algorithms with n-gram character features". In: (2012).

[32] Isabel Briggs Myers and Peter B Myers. *Gifts differing: Understanding personality type*. Nicholas Brealey, 2010.

[33] Clifford Nass and Kwan Min Lee. "Does computer-generated speech manifest personality? An experimental test of similarity-attraction". In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 2000, pp. 329–336.

[34] Mª Ángeles Oviedo-García et al. "Metric proposal for customer engagement in Facebook". In: *Journal of research in interactive marketing* (2014).

[35] Wendy Patton and Mary McMahon. *Career development and systems theory: Connecting theory and practice*. Vol. 2. Springer, 2014.

[36] Marco Pennacchiotti and Ana-Maria Popescu. "A machine learning approach to twitter user classification". In: *Fifth international AAAI conference on weblogs and social media*. 2011.

[37] James W Pennebaker and Laura A King. "Linguistic styles: Language use as an individual difference." In: *Journal of personality and social psychology* 77.6 (1999).

[38] Barbara Plank and Dirk Hovy. "Personality Traits on Twitter—Or—How to Get 1,500 Personality Tests in a Week". In: *The 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA), EMNLP 2015*. 2015.

[39] Daniele Quercia et al. "Our twitter profiles, our selves: Predicting personality with twitter". In: *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE. 2011, pp. 180–185.

[40] Daniel M Romero et al. "Influence and passivity in social media". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2011, pp. 18–33.

[41] Xianzhi Ruan, Steven Wilson, and Rada Mihalcea. "Finding optimists and pessimists on twitter". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2016, pp. 320–325.

[42] Arthur L Samuel. "Some studies in machine learning using the game of checkers". In: *IBM Journal of research and development* 3.3 (1959), pp. 210–229.

[43] H Andrew Schwartz et al. "Extracting human temporal orientation from Facebook language". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2015, pp. 409–419.

[44] H Andrew Schwartz et al. "Personality, gender, and age in the language of social media: The open-vocabulary approach". In: *PloS one* 8.9 (2013), e73791.

[45] Youngseo Son et al. "Recognizing counterfactual thinking in social media texts". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2017, pp. 654–658.

[46] Christian Torrero, Carlo Caprini, and Daniele Miorandi. "A Wikipedia-based approach to profiling activities on social media". In: *arXiv preprint arXiv:1804.02245* (2018).

[47] Ben Verhoeven, Walter Daelemans, and Barbara Plank. "Twisty: a multilingual twitter stylometry corpus for gender and personality profiling". In: *Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016)/Calzolari, Nicoletta [edit.]; et al.* 2016, pp. 1–6.

[48] Wu Youyou, Michal Kosinski, and David Stillwell. "Computer-based personality judgments are more accurate than those made by humans". In: *Proceedings of the National Academy of Sciences* 112.4 (2015), pp. 1036–1040.

[49] Mohammadzaman Zamani, Anneke Buffone, and H Andrew Schwartz. "Predicting Human Trustfulness from Facebook Language". In: *arXiv preprint arXiv:1808.05668* (2018).