Information Retrieval Journal
https://doi.org/10.1007/s10791-018-9338-x

SOCIAL MEDIA FOR PERSONALIZATION AND SEARCH

Determining the interests of social media users: two approaches

Nacéra Bennacer Seghouani[1] · Coriane Nana Jipmo[1] · Gianluca Quercini[1]

Received: 11 November 2017 / Accepted: 28 June 2018

Abstract
Although social media platforms serve diverse purposes, from social and professional networking to photo sharing and blogging, people frequently use them to share the thoughts and opinions and most importantly, their interests (e.g., politics, economy, sports). Understanding the interests of social media users is key to many applications that need to characterize them to recommend some services and find other individuals with similar interests. In this paper, we propose two approaches to the automatic determination of the interests of social media users. The first, that we named Frisk, is an unsupervised multilingual approach that determines the interests of a user from the explicit meaning of the words that occur in the user's posts. The second, that we termed Ascertain, is a supervised approach that resorts to the hidden dimensions of the words that several studies indicated to be capable of revealing some of the psychological processes and personality traits of a person. In our evaluation, that we performed on two datasets obtained from Twitter, we show that Frisk is capable of inferring the interests in a multilingual context with good accuracy and that the psychological dimensions used by Ascertain are also good predictors of a user's interests.

Keywords Interest classification · Twitter user profile · Personality

1 Introduction

The popularity of online social media platforms has been steadily growing over the years. According to a recent survey released by Pew Research, two-thirds of American adults use at least one social media platform compared to only 7% back in 2005 (Perrin 2015). While the social media demographics is markedly skewed towards younger generations, social

✉ Gianluca Quercini
gianluca.quercini@lri.fr

Nacéra Bennacer Seghouani
nacera.bennacer@lri.fr

Coriane Nana Jipmo
coriane.nanajipmo@lri.fr

[1] Laboratoire de Recherche en Informatique LRI, CentraleSupélec, University of Paris-Saclay, Paris-Saclay, France

media users are undoubtedly a representative sample of the adult population, also considering that the past few years have seen a significant increase in the usage of older adults. This, coupled with the public availability of the content published by social media users, has fueled a lot of research aimed at better understanding how people interact and behave. Although today's social media platforms serve diverse purposes, from social (e.g., Facebook, Twitter) and professional networking (e.g, Linkedin) to photo sharing (e.g., Flickr) and blogging (e.g., Livejournal), one common denominator is that people frequently use them to share their thoughts and opinions, and talk of what they like—and what they are interested in. Several studies have shown that the textual posts (i.e., text published) by social media users is a good indicator of their *interests* (Ding and Jiang 2014; Michelson and Macskassy 2010; Raghuram et al. 2016; Spasojevic et al. 2014; Vu and Perez 2013; Wang et al. 2013; Wen and Lin 2011; Weng et al. 2010; Xu et al. 2011; Zarrinkalam et al. 2015). After all, it comes as no surprise that those who are primarily interested in politics will mostly write about elections, legislation and major political events.

Understanding the interests of social media users, which is the focus of this paper, is key to many applications that need to characterize the individuals to recommend them some services (Pennacchiotti et al. 2012), find other individuals with similar interests (Wang et al. 2011), and even predict their future interests (Bao et al. 2013). More formally, the problem of determining the interests of a user u can be defined as follows:

Input: $\mathcal{I} = \{I_1, I_2, \ldots, I_n\}$; $\mathcal{T}_u = \{t_1, t_2, \ldots, t_m\}$
Output: $r_u : \mathcal{I} \rightarrow \mathbb{R}$
where:

- $\mathcal{I}$ is a set of n interests $I_1, \ldots, I_n$.
- $\mathcal{T}_u$ is a set of m textual posts authored by user u.
- $r_u : \mathcal{I} \rightarrow \mathbb{R}$ is the *relevance score function* that maps each interest to a numeric score.

Ideally, the interest with the highest relevance score is the dominant interest of user u. The use of a relevance score allows to rank the interests, which is particularly important because a social media user might have several.

Not surprisingly, a lot of studies focused on the problem of determining the interests of individuals based on what they write. The existing approaches can be categorized along different axes, based on the data that they exploit (content written by the individuals themselves vs. their friends) and the methods that they use (supervised vs. unsupervised). Supervised methods generally use text classifiers that assign users to one class or a probability distribution over a set of classes that represent their interests (Raghuram et al. 2016; Spasojevic et al. 2014). These methods have the merit of providing a clear categorization of the interests, because a class, that has a specific label (e.g., *politics*, *economics*), corresponds to an interest. On the other side, they have two major shortcomings. First, they need a training set, where users are manually categorized by their interests based on a visual inspection of what they write, which is a time-consuming task. Second, they are language-dependent because a classifier trained on posts written in English cannot be used to predict the interests of a person who writes in French. Unsupervised approaches (Michelson and Macskassy 2010; Vu and Perez 2013; Wang et al. 2013; Wen and Lin 2011; Weng et al. 2010; Xu et al. 2011; Zarrinkalam et al. 2015) describe the interests through bags of words (e.g., *phone*, *developers*, *apps*) or sets of concepts derived from Wikipedia that are often difficult to interpret. The advantage is that they are independent of the language and they need no training data.

In a previous paper of ours, we described Frisk (Find twitteR InterestS via wiKipedia), an approach that infers the interests of a user from the explicit meaning of the words that occur in the user's posts (Jipmo et al. 2017). The rationale of Frisk is that users interested in a topic will use a lot of words related to that topic. More precisely, Frisk is an unsupervised multilingual approach that models the tweets (i.e., textual posts in Twitter) of a user and the interests (e.g., *politics*, *sports*) as bags of articles and categories of Wikipedia respectively, and ranks the interests by their relevance score, measured in terms of the graph distance between the articles and the categories. Unlike the existing unsupervised approaches (Michelson and Macskassy 2010; Vu and Perez 2013; Wang et al. 2013; Wen and Lin 2011; Weng et al. 2010; Xu et al. 2011; Zarrinkalam et al. 2015), Frisk does not output the interests as bag-of-words, or bag-of-concepts, but it gives a precise categorization of the interests, in the same way as supervised approaches do, but with the advantage of being independent of the language and without needing training data.

In this paper, we introduce a new approach for determining the interests of social media users that we name Ascertain (AnalySis Correlation pERsonality Traits And INterests) and we compare it against Frisk. As opposed to Frisk, Ascertain resorts to the hidden dimensions of the words that several studies indicated to be capable of revealing some of the psychological processes and personality traits (based on the Big Five model) of a person (Schwartz et al. 2013; Tausczik and Pennebaker 2010). The hypothesis behind Ascertain is that the personality traits of a user are a good predictor of the interests of that user, supported by the fact that positive correlations have been found between personality and vocational (i.e., professional) interests (Gottfredson et al. 1993) and social behavior (Furnham and Heaven 1999). Ascertain is a supervised approach that represents each user in a low-dimensional space consisting of features that reflect the psychological dimensions of the user. These features are derived from the textual posts of a user by using a software called Receptiviti that is based on LIWC (Linguistic Inquiry and Word Count), a well-known text processing tool (Pennebaker et al. 2015). The interests of a user are determined by learning a linear regression classifier on these features. Unlike Frisk, we never described Ascertain in any previous paper.

In summary, we claim the following contributions:

- We analyze a real dataset obtained from Twitter to uncover the correlation between the personality traits and the interests of a user.
- We propose an approach called Ascertain that determines the interests of a user from its personality traits and psychological dimensions, as computed by Receptiviti from the text written by that user. To the best of our knowledge, this is the first approach of such nature.
- We study the contribution of three different sets of Receptiviti dimensions to the prediction of user interests.
- We compare Ascertain against our previous approach Frisk that is capable of inferring interests in a multilingual context with good accuracy. The interesting part here is the comparison between an approach—Frisk—that uses the explicit meaning of the words to infer interests and another approach—Ascertain—that uses hidden dimensions that the words can reveal.
- We thoroughly evaluate both approaches on two real datasets obtained from Twitter and we compared them with two existing approaches, one based on text classification (Raghuram et al. 2016) and the other based on LDA (Weng et al. 2010). Although the evaluation is confined to Twitter due to the large availability of data, the same consid-

erations can be generalized to any social media platform that let its users post textual content.

The remainder of the paper is organized as follows. After a thorough review of the literature (Sect. 2), we describe the two datasets that we use for the evaluation of the two approaches and the methodology that we followed to obtain them (Sect. 3). The approaches Frisk and Ascertain are presented in Sects. 4 and 5 respectively; in both sections, the necessary terminology and notation are introduced before detailing the approaches. The experimental results are discussed in Sect. 6 followed by concluding remarks and perspectives in Sect. 7.

2 Related work

The scientific literature describes many approaches to infer the interests of social media users based on what they, or their friends, write (Ding and Jiang 2014; Michelson and Macskassy 2010; Raghuram et al. 2016; Spasojevic et al. 2014; Vu and Perez 2013; Wang et al. 2013; Wen and Lin 2011; Weng et al. 2010; Xu et al. 2011; Zarrinkalam et al. 2015), the tags that they use to organize their favorite resources (e.g., bookmarks of Web pages) (Li et al. 2008; Wang et al. 2011) and the celebrities that they like or follow (Bhattacharya et al. 2014; He et al. 2015). In Sriram et al. (2010), the authors have proposed to use a small set of domain-specific features extracted from the authors profile and text to effectively classify the text to a predefined set of generic classes such as News, Events, Opinions, Deals, and Private Messages. Our approaches differ in that the classes represent interests of users instead of generic categories.

We focus here on the approaches that are based on the textual content posted in Twitter. Since few Twitter users mention their interests in their biographies (Ding and Jiang 2014), researchers usually turn their attention to the tweets.

Supervised machine learning techniques trained on textual and profile features, such as gender and location, proved to be effective (Raghuram et al. 2016; Spasojevic et al. 2014), but they are language-dependent and need a training set that is difficult to obtain.

As for the unsupervised approaches, Vu and Perez use regular expressions to extract key-phrases from the tweets and rank them by their frequency based on popular measures such as tf-idf or TextRank (Vu and Perez 2013). Many researchers resort to a technique known as *Latent Dirichlet Allocation* (LDA) (Wang et al. 2013; Weng et al. 2010; Xu et al. 2011), a probabilistic topic model for uncovering the topics of a collection of textual documents (Blei et al. 2003).

In essence, LDA sees the textual content of a document as a mixture of topics (interests, in our context), according to a certain probability distribution (e.g., 60% *politics*, 40% *economics*), and each topic is a probability distribution over words (e.g., 40% *election*, 30% *campaign*, 30% *Congress* for the topic *politics*). A document is the result of a generative process that randomly chooses a topic, according to the given probability distribution, and samples a word from that topic. While Weng et *al.* apply LDA as is to the tweets (Weng et al. 2010), Xu et al. describe a variation of LDA that filters out noisy tweets that do not necessarily reflect the interests of the user (e.g., tweets about everyday activities) (Xu et al. 2011). The main shortcoming of all these approaches is that they describe an interest as a bag of words, or a probability distribution over words, without giving the exact categorization of the interest (e.g.,

politics). Moreover, our experiments show that LDA does not perform well in a multilingual context.

Among the approaches that use Wikipedia, Zarrinkalam et al. model the relationships among the concepts, represented as Wikipedia articles, mentioned in a user's tweets in a certain time interval and identify the interests as clusters of concepts (Zarrinkalam et al. 2015). These clusters are unnamed and often hard to interpret. Michelson and Macskassy describe the interests of a Twitter user as a bag of Wikipedia categories (e.g., *Football in England*) determined from the named entities (e.g., *Arsenal*, *Walcott*) mentioned in the tweets (Michelson and Macskassy 2010). The approach described in Kapanipathi et al. (2014) provides a framework for the identification of broader user interests based on their tweets. The interests are represented as a hierarchy of Wikipedia categories called *Hierarchical Interest Graph* (HIG). In the same way as the approach described in Zarrinkalam et al. (2015), they only use the named entities to identify Primitive Interests from a users tweets. The evaluation has been conducted on 37 Twitter users.

As opposed to that, Frisk and Ascertain do not only consider named entities but also generic words, they categorize precisely the interests (instead of representing them as bag of concepts) and the evaluation is conducted on datasets including up to 1347 users.

There has been growing interest on characterizing users especially their personality (Rentfrow and Gosling 2003; Rentfrow et al. 2011; Rawlings and Ciancarelli 1997; Cantador et al. 2013; Odic et al. 2013; Schwartz et al. 2013). The personality is defined as the set of habitual behaviors, cognitions and emotional patterns that characterize a person. The Big Five personality traits (openness, conscientiousness, extroversion, agreeableness, neuroticism) is a model based on common language descriptors to describe the human personality. These traits of personality are correlated with the lexical common language and the personality characteristics that are most important in people's lives could be a part of their language.

Some works in the literature (Rentfrow and Gosling 2003; Rawlings and Ciancarelli 1997) focused on studying the correlation between music preferences and Big Five traits. These studies have shown that music preferences mostly correlate with extroversion and openness. A preliminary study (Cantador et al. 2013) evidences the relations between personality types and user preferences in multiple entertainment domains, namely movies, TV shows, music, and books. They extracted personality-based user stereotypes from 53,226 Facebook profiles composed of both personality scores from the Big Five model, and explicit interests about 16 genres in each of the above domains. Odic et al. (2013) presented a preliminary work on understanding the impact of personality on the emotion induction during the consumption of movies.

Besides these research work, the study that motivated our approach Ascertain (Schwartz et al. 2013) describes the correlations between the written texts and the personality based on the Big Five model. Our objective is to go one step further by investigating the correlation between personality and interests; based on these correlations, we trained a supervised model to predict users' interests (not just in the domain of music and movies). Moreover, one important contribution of this paper is the comparison between an approach, like Frisk, that is based on the explicit semantics of the words and another approach, Ascertain, that uses the psychological dimensions hidden in the words.

Table 1 Statistics on MULTIDS

Interest	#users					Tweets	
	en	fr	it	es	Total	#tweets	#avg(#tweets)
Politics	92	56	69	57	**274**	739,499	2698.9
Economy	93	58	57	68	**276**	596,331	2160.6
Games	85	58	47	58	**248**	625,108	2520.6
Gastronomy	89	53	73	50	**265**	492,197	1857.3
Sports	88	66	67	63	**284**	822,597	2896.5
Grand total	447	291	313	296	**1347**	**3,275,732**	**12,133.9**

Best values are highlighted in bold

3 Dataset

We collected two datasets from Twitter; one multilingual (denoted MULTIDS) that includes four disjoint sets that contain users tweeting in English (en), French (fr), Italian (it) and Spanish (es) respectively; and one monolingual (denoted MONODS) that includes users who only tweet in English. To obtain the two datasets, we followed a two-step methodology that consists in (i) collecting a set of Twitter users categorized by interest and (ii) retrieving the most recent tweets (up to 5000) of each user.

In order to collect the set of users, we submitted queries to the Twitter search engine, using keywords that are representative of five selected interests—*Politics*, *Economy*, *Games*, *Gastronomy* and *Sports*—and a sixth interest, only for the dataset MONODS—*tourism*. For each interest (between parentheses in the following list) we obtained the following keywords from the category hierarchy provided by Google AdWords[1]: politics, government, campaigns, elections (*Politics*); economy, finance (*Economy*); game, board games, video games (*Games*); sports, basketball, baseball, bowling, football (*Sports*); drink, food, restaurant (*Gastronomy*); tourism, travel, museum (*tourism*). We manually translated these keywords to French, Italian and Spanish. Finally, we filtered out the users who did not have any of the five selected interests by reading the tweets of all users, a long manual process needed to make sure that the ground truth was reliable. This step is necessary because we observed that the Twitter search engine might return users that are not relevant to the selected interests even if they somehow match the given keywords.

At the end of the collection, the dataset MULTIDS consists of 1347 users, evenly distributed across four disjoint subsets *en*, *fr*, *it*, *es* that contain users who tweet in English, French, Italian and Spanish respectively. distributed uniformly over the four languages, and more than 3 million tweets (Table 1). The dataset MONODS consists of 1446 users who tweet in English, distributed over six interests, and more than 600,000 tweets (Table 2).

[1] developers.google.com/adwords/api/docs/appendix/productsservicesdevelopers.google.com/adwords/api/docs/appendix/productsservices.

Table 2 Statistics on MonoDS

Interest	#users(en)	#tweets	#avg(#tweets)
Politics	250	111,382	445.5
Economy	202	83,023	411
Games	235	105,595	449.3
Gastronomy	219	95,974	438.2
Sports	286	136,094	475.9
Tourism	254	115,622	455.2
Grand total	**1446**	**647,690**	**2675.21**

Best values are highlighted in bold

4 Frisk: Find twitteR interestS via wiKipedia

Friskranks the interests of a Twitter user by relevance based on the tweets authored by that user. The rationale is that the words occurring in the tweets (e.g., *campaign*, *elections*, *player*, *game*) are important clues as to what the user likes (e.g., *politics*, *sports*); the more the words related to an interest, the higher the relevance of that interest.

The key to the approach is the computation of a relevance score of an interest from the words of the tweets by using a Wikipedia-based representation. We observe that a Wikipedia article (e.g., *Political campaign*) identifies a specific meaning of a word (e.g., *campaign*) and thus a set of tweets can be represented as the bag (i.e., collection) of the Wikipedia articles associated to its words. At the same time, a Wikipedia article is organized into a hierarchy of categories (e.g., *Category:Political activism*, *Category:Politics*) that identify the broad domain (e.g., *politics*) of that article and, as such, the interest of a person who reads the article. An interest can therefore be described as a bag of Wikipedia categories. As a result, the words occurring in the tweets and the interests are both described as nodes (articles and categories, respectively) of the Wikipedia graph. Frisk computes the relevance score of an interest to a Twitter user in terms of the graph distance between the articles associated to the tweets of that user and one of the categories representing the interest.

Friskconsists of two major steps:

1. *Tweets analysis* In this step, Frisk determines the set of Wikipedia articles that describe the tweets of u. This is achieved in two phases:
   a *Preprocessing* In this phase, Frisk removes from the tweets the text that unlikely to be useful (e.g., stop words) to characterize the interests of u. The output is a bag-of-words $\mathcal{BOW}_u$.
   b *Bag-of-articles representation*. In this phase, Frisk associates, whenever possible, a Wikipedia article to each word in $\mathcal{BOW}_u$. The output is a bag-of-articles $\mathcal{BOA}_u$, $|\mathcal{BOA}_u| \leq |\mathcal{BOW}_u|$.

2. *Interests determination* In this step, Frisk uses the bag-of-articles $\mathcal{BOA}_u$ to determine the interests of u, in two phases:
   a *Relevance score computation* Frisk computes the relevance score function r.
   b *Interest ranking* Frisk sorts the interests by relevance score in descending order. Ideally, the top-ranked interest is the dominant interest of u.

In the remainder of this section, we describe each step in greater detail. First, we introduce the basic notion of Wikipedia that are necessary to understand the approach.

4.1 Wikipedia

Friskuses Wikipedia, the largest online multilingual encyclopedia with 284 active language editions, to represent both the tweets of a user and the interests. Any Wikipedia edition in a language α consists of a set of interlinked pages. A *page* has one of four types: article, disambiguation, redirect or category. An *article* discusses a specific topic and consists of a *title* (e.g., *Paris*), a textual content, *links* to related articles (e.g., *Eiffel Tower* and *Louvre*) and *cross-language links* to articles that cover the same topic (e.g., *Parigi*) in other language editions (e.g., the Italian Wikipedia). A *disambiguation page* with title t (e.g., *Campaign*) contains links to Wikipedia articles that are possible interpretations of t (e.g., *Advertising campaign*, *Political campaign*). A *redirect page*, or simply *redirect*, with title t (e.g., *Paris, France*) has no content itself and provides a link to another article whose title is an alias of t (e.g., *Paris*). Henceforth, we will use the title to refer to a page. Each page (e.g., the article *Political campaign*), except redirects, is included in one or more *categories* (e.g., *Category:Political campaigns*). The motivation for categories is to help readers browse the pages, and to this extent they are organized in a hierarchy, as each may branch into *subcategories*, as well as possibly being included in one or more categories.

In this paper we model Wikipedia as a directed graph $\mathcal{W} = (PA, LI)$, where the nodes in PA correspond to the pages and the edges in LI are pairs of nodes connected by a link. Unless otherwise specified, we will treat the terms node and Wikipedia page as synonyms. A node p_α belongs to a Wikipedia edition in a specific language α and has one of the four types specified above. A cross-language link (p_α, p_β) (or simply *crosslink*) connects two pages p_α and p_β that cover the same topic in two different language editions α and β. The meaning of an *intra-language* link (or simply, link), connecting a page p_α to a page q_α within the same language edition, depends on the types of p_α and q_α. If p_α is a disambiguation page, q_α corresponds to one of the possible interpretations of p_α; if p_α is a redirect page, q_α is the page to which p_α redirects; if p_α is an article and q_α is a category, p_α is contained in the category q_α; finally, if both p_α and q_α are categories, p_α is contained in q_α (q_α is *parent* of p_α, or, alternatively, p_α is a *subcategory* of q_α). We note that if there is a path from an article a_α to a category c_α with length k (i.e., k links), the article a_α is contained in the category c_α, though not directly, and, as such, c_α is called a *k-ascendant* of a_α.

4.2 Tweets analysis

The purpose of this step is to represent the tweets of user u as a set of Wikipedia articles. This is achieved by first obtaining a representation of the tweets as a bag-of-words.

4.2.1 Preprocessing

Tweets in $\mathcal{T}_u$ are preprocessed to remove the stop words (e.g., words that occur frequently in a text without adding any meaning, such as articles and prepositions) in English, French, Italian and Spanish, as given by the NLTK corpus. Numbers, special characters (e.g., /, \, #) and URLs are also removed. Mentions (i.e., references to Twitter users intended to be the recipient of a tweet) are removed from the tweets because they will almost never map to any Wikipedia article, with very few exceptions (e.g., *@neo4j*). We keep *hashtags* (i.e.,

keywords preceded by the character "#" used to specify the topic of a tweet) while removing the character "#", as they might map to Wikipedia articles, although we found that it does not occur frequently, due to the fact that an hashtag might be composed of multiple words without spaces (e.g., "#androidgames"). The output of this step is a bag of words representation $\mathcal{BOW}_u$ of the tweets $\mathcal{T}_u$. As an example, consider the following tweet :

@Jules Going to the store without a budget #creditchat

where "@Jules" is a mention, "#creditchat" is a hashtag and "to", "the", "without" and "a" are stop words. The bag-of-words associated with this tweet is as follows :

{store;budget;creditchat}

4.2.2 Bag-of-articles representation

When it comes to transforming the bag of words $\mathcal{BOW}_u$ into a bag of articles $\mathcal{BOA}_u$, the challenge is that natural language is inherently ambiguous. Thus, a word w (e.g., *gender*) might have several *meanings* (e.g., distinction between male and female, or grammatical gender), that are described by a set A_w of different Wikipedia articles, also referred to as the *interpretations* of w. When a person uses a word in a text, that word assumes only the meaning that corresponds to the message that the person intends to communicate. Therefore, we need a method that selects one interpretation of any given word w, a process that is known as *word disambiguation* and is a widely studied research topic (Navigli 2009).

The additional challenge in our context is that the users of Twitter, and social media in general, are spread across the globe and write in different languages. We cannot even assume that the tweets of a single person are only written in her native language, because she might speak many and use all of them on a regular basis to communicate her thoughts. To the best of our knowledge, the most prominent Wikipedia-based approaches to word sense disambiguation in a multilingual context are TagMe (Ferragina and Scaiella 2010) and Babelfy (Moro et al. 2014). We decided to use the former for three reasons:

1. It is specifically conceived to process texts that are "short and poorly composed" (Ferragina and Scaiella 2010), which makes it the ideal tool to disambiguate words in tweets.
2. It provides an easy-to-use RESTful API with no limits on the number of daily queries, unlike Babelfy.
3. Most importantly, it always associates words to Wikipedia articles, unlike Babelfy that often maps words to concepts obtained from other resources (e.g., WordNet).

The only limit of TagMe is that it is not language agnostic in that it needs to know the language of a text before processing it. For this reason, besides using TagMe, we also propose a language agnostic heuristic to obtain $\mathcal{BOA}_u$. More precisely, we propose two versions of FRISK, namely FRISKTM and FRISKLA; the former creates $\mathcal{BOA}_u$ by using TagMe, the latter by using a language-agnostic heuristic. Both approaches are described in the remainder of this section.

FRISKTM creates $\mathcal{BOA}_u$ by using TagMe. Taking in the preprocessed tweets of user u, TagMe assigns a score to each interpretation of a word that reflects the probability that that interpretation is the right one for the word, based on the other words that occur in the text. More specifically, the score assigned to an interpretation a of a word w is obtained by summing all the votes that a receives from all the words in the text. The vote $vote_v(a)$ of a

word v to the interpretation a is obtained as the weighted average relatedness between each interpretation b of v and the interpretation a, as follows:

$$vote_v(a) = \frac{\sum_{b \in A_v} rel(b, a) \cdot Pr(b|v)}{|A_v|}$$

where:

- A_v is the set of all interpretations of the word v.
- $rel(b, a)$ is the relatedness between the interpretations a and b, calculated with the semantic relatedness measure WLM (Witten and Milne 2008).
- $Pr(b|v)$ is the weight assigned to the relatedness $rel(b, a)$ and is the probability that b is the right interpretation of the word v (most common interpretation).

FriskLA creates $\mathcal{BOA}_u$ by using a language-agnostic heuristic. This heuristic does not make any assumption as to the language of the tweets, it looks for the interpretations of a word w across all Wikipedia language editions. In order to address the problem of ambiguity, Based on the Wikipedia guidelines on disambiguation, the Wikipedia article that has the word w as its title is usually the *default* (i.e., most common) interpretation of w and, as such, the correct interpretation in many cases. This stems from the fact that for an ambiguous word (e.g., *election*) the article that corresponds to the default interpretation (e.g., the one describing the political decision-making process) is chosen by millions of Wikipedia users that agree that the word is normally used with that meaning. For this reason, the heuristic always selects the default interpretation of w, if it exists, as the correct one.

If the word w (e.g., *campaign*) is ambiguous and the Wikipedia users cannot find any agreement as to its default interpretation, a disambiguation page exists having w as its title and listing all the possible interpretations of w (e.g., *Political campaign*, *Advertising campaign*). In this case, the heuristic selects the interpretation (Wikipedia article) with the highest indegree, based on the observation that the number of Wikipedia articles that link to an interpretation is an indication of the importance, or *popularity*, of that interpretation.

The choice of consistently assigning a word either the default or the most popular interpretation has the obvious downside of penalizing less common interpretations. For instance, the word *java* will always be linked to the Wikipedia article that describes the Indonesian island. However, someone who tweets and shows its interest in software development, might use other words (e.g., *Python*, *.NET*) that will be linked to the correct interpretations, leading to the correct prediction of its interest, even if some words are incorrectly disambiguated.

Algorithm 1 Calculate $\mathcal{BOA}_u$

```
1: function COMPUTEBOA(BOW_u, W): BOA_u
2:     BOA_u ← ∅
3:     for each w ∈ BOW_u do
4:         A_w = ∅
5:         A = getArticlesByTitle(w, W)
6:         for each a ∈ A do
7:             if isRedirect(a) then a ← resolveRedirect(a, W)
8:             end if
9:             if isDisambiguationPage(a) then
10:                I ← getInterpretations(a, W)
11:                a ← getMostPopular(I, W)
12:            end if
13:            if language(a) ≠ "English" then a ← geEnglishArticle(a, W)
14:            end if
15:            if |categories(a)| ≥ θ then A_w ← A_w ∪ {a}
16:            end if
17:        end for
18:        BOA_u ← BOA_u ∪ {getMostPopular(A_w, W)}
19:    end for
20: end function
```

The function that computes $\mathcal{BOA}_u$ (Algorithm 1) iterates over each word w in $\mathcal{BOW}_u$ and obtains the set A of articles that have w as their title across all language editions of Wikipedia $\mathcal{W}$ (Line 5). For each article $a \in A$, if a is a redirect page, the article that is the target of the redirection is obtained (Line 7). If a is a disambiguation page (i.e., there is no default interpretation of the word w), the function gets all the interpretations (i.e., all the articles to which the disambiguation page links) and selects the most popular (Line 11); otherwise, a is the default interpretation and is selected.

In the example of Fig. 1, the word "election" is looked for across three Wikipedia language versions (English, French and Italian). In each of the three versions there exist one page whose title is *Election*:

- In the English version, the page is an article that is the default interpretation. Note that $p : 1236$ is the value of the popularity (i.e., number of incoming links) of this interpretation in the English Wikipedia.
- In the French Wikipedia, the page redirects to a disambiguation page titled *ÉElection (homonymie)* that links to three articles (possible interpretations), respectively titled *ÉElection*, *Election 2* and *Election (film)*. The first one is selected due to its higher popularity ($p : 575$) in the French Wikipedia.
- In the Italian Wikipedia, the page is a disambiguation that links to two articles, respectively titled *Election (film 1999)* and *Election (film 2005)*. The first one is selected because of its higher popularity ($p : 42$) in the Italian Wikipedia.

The example highlights the fact that the heuristic selects, as a possible interpretation of the word w, an article in each Wikipedia edition; as a result, the word w may be associated to two or more (possibly) conflicting interpretations. In order to select only one interpretation of w, the English version of each selected interpretation is obtained by using the existing cross-language links (Line 13). Therefore, a set of possible interpretations is obtained that belong to a single Wikipedia edition. In the example of Fig. 1, the interpretation

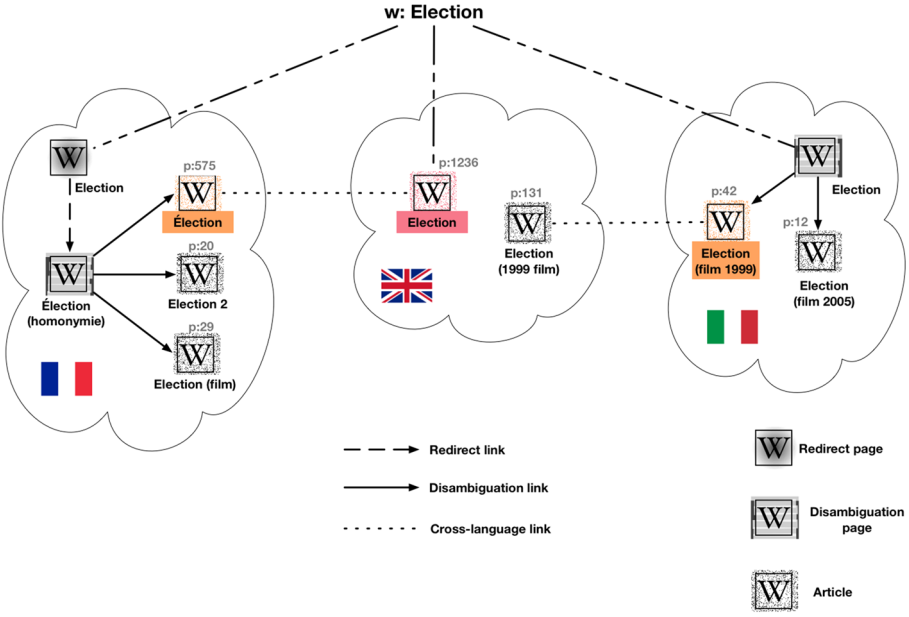


Fig. 1 The language agnostic heuristic

selected in the French Wikipedia (resp., the Italian Wikipedia) is mapped to the article titled *Election* (resp., *Election (1999 film)*) in the English Wikipedia. Finally, the interpretation with the highest popularity (the one titled *Election* in the example) is selected as the interpretation of the word w (Line 18). Note that the interpretations that are in less than θ Wikipedia categories are ignored (Line 15). We experimentally set $\theta = 5$ after several tests. The bag-of-articles of the bag-of-words {store, budget, creditchat} obtained in Sect. 4.2.1 is as follows:

$$\{\text{Retail, Budget}\}$$

where the word "store" is associated to the article *Retail*, the word "budget" is associated to the article *Budget* and the word "creditchat" has no interpretation.

4.3 Interest ranking

In order to rank the interests by relevance to the user u, Friskcomputes a relevance score $r_u(i)$ for each interest $i \in \mathcal{I}$ by using the bag-of-articles $\mathcal{BOA}_u$ describing the tweets of u and the Wikipedia categories. Suppose that $\mathcal{BOA}_u$ contains the article titled *Goalkeeper* that reflects a possible interest of the user in *sports*. This article belongs to a category named *Sports terminology* that, in turn, is contained into the categories named *Sports* and *Sports rules and regulations*. If we consider the distance between two nodes in the Wikipedia graph as the length of the shortest path leading from one to the other, the article *Goalkeeper* is certainly closer to any of the sports categories than a, say, category that includes articles related to politics. In other words, if we describe an interest i as a *bag-of-categories* $\mathcal{BOC}_i$, we can measure the relevance of an article a to an interest i based on the graph distance between a and any of the categories in $\mathcal{BOC}_i$; the shorter the distance, the

higher the relevance. More formally, the relevance $r(a, i)$ of an article a to an interest i can be expressed as follows:

$$r(a, i) = \frac{1}{\min_{c \in \mathcal{BOC}_i} dist(a, c)} \tag{1}$$

The relevance score $r_u(i)$ of an interest is obtained as the sum of the relevance scores of all articles $a \in \mathcal{BOA}_u$ to the interest i, as follows:

$$r_u(i) = \sum_{a \in \mathcal{BOA}_u} r(a, i) \tag{2}$$

In order to calculate the distance from an article a to all the categories in $\mathcal{BOC}_i$, we need to perform a visit of the Wikipedia graph starting from a and following the links that lead to its parent and ancestor categories. Since this visit may be costly, depending on the distance of the categories from a, we cap the visit to the k-ancestors of a. Based on our experiments, we found that an acceptable value is $k = 3$; for higher values, indeed, we are likely to find categories that are generically related to a.

Finally, there are multiple options to obtain the bag-of-categories $\mathcal{BOC}_i$ associated to an interest i. One possibility is to select the category named after the interest (e.g., *Sports*)—the root category—and get all its subcategories, possibly down to a predetermined depth. This approach would be a good choice if the subgraph induced by the categories was a tree representing a well-organized taxonomy, which is not the case. In fact, the subgraph of the categories is not even a DAG, let alone a tree, and some subcategories at a certain depth might have nothing to do with the root category. For instance, as of the time of writing, the category named *Firearm laws* is a subcategory of *Sports* at depth 3 (*Sports* → *Politics and sports* → *Gun politics* → *Firearm laws*). As a result, this approach would end up adding to $\mathcal{BOC}_i$ all the categories relevant to an interest, all the while including too many that are not. Instead, we add into $\mathcal{BOC}_i$ all categories (e.g,, *Team Sports*, *Politics awards*, *Economy of France*) that have the name of the interest (or its stem) (e.g., *sport*, *politics*, *economy*) in their titles, which is a strong indication as to their relevance to the respective interests. Although many categories that are relevant to an interest i will be left out of $\mathcal{BOC}_i$ (e.g., *Football in Italy*), Frisk will still be able to find some of the categories in $\mathcal{BOC}_i$ when crawling the categories of an article a up to its third level of ancestors.

For these reasons, we opted for a simpler, yet more effective, method that consists in representing i (e.g., *sports*) as the bag of Wikipedia categories $\mathcal{BOC}_i$ whose titles contain the name of the interest (e.g., *Category:Sports by year*, *Category:Sport in France*).

As a final remark, we note that a possible improvement of the ranking procedure would be to count the number of times a category is visited when Frisk crawls the category of each article in the bag-of-articles. This would weaken the impact on the final relevance score of those categories that are highly relevant to only one article. This improvement is left for future work.

5 Ascertain: AnalySis Correlation pERsonality Traits And INterests

Similarly to Frisk, Ascertainis an approach that predicts the interests of a user based on the tweets or, for that matters, any text, authored by that user. However, Frisk does so by exploiting the *meaning* of the words in the tweets, under the sound hypothesis that users interested in a topic will use a lot of words related to that topic. As opposed to this,

ASCERTAIN resorts to the *hidden* dimensions of the words that several studies indicated to be capable of revealing some of the psychological processes and personality traits of a person (Schwartz et al. 2013; Tausczik and Pennebaker 2010).

Although positive correlations have been found between personality and professional vocation (Gottfredson et al. 1993) and social behavior (Furnham and Heaven 1999), no previous work we are aware of has shown a correlation between personality traits and recreational interests, which is the innovative part of this work.

Given a set of interests $\mathcal{I}$ and the set of tweets $\mathcal{T}_u$ of a user u, ASCERTAIN determines the interests of user u through the following steps:

- Transform the tweets in $\mathcal{T}_u$ into a feature vector, where each feature represents a psychology dimension, by using a tool named Receptiviti (cf. Sect. 5.1);
- Select the features that are likely to be good predictors of the interests.
- Compute the relevance score $r_u(i)$ of each interest i to the user u as the probability $Pr(i|u)$ by using a logistic regression classifier.

We detail these steps in Sect. 5.3, but first we introduce the necessary terminology (Sect. 5.1) and we present the results of the exploration of our data through principal component analysis that support the hypothesis that the personality traits actually correlate with some of the interests (Sect. 5.2).

5.1 Background and terminology

The *Encyclopedia of Psychology* defines personality as the "individual differences in characteristic patterns of thinking, feeling and behaving" (Kazdin 2000). Although the differences among individuals can be numerous, early investigations using factor analysis showed that only five factors—commonly referred to as *personality traits* (Eysenck 2012)—are sufficient to describe the personality of a person (Thurstone 1934).

These five factors, which recur in several studies, have become known as the **Big Five Model** (short, Big5), a widely accepted framework for the description of personality (Digman 1990; Goldberg 1990). More precisely, the five traits are labeled as follows:

1. Openness. Being open to new ideas and experiences.
2. Conscientiousness. Being well-organized and self-disciplined.
3. Extraversion. Being positive and outgoing.
4. Agreeableness. Being compassionate and cooperative.
5. Neuroticism. Being negative and prone to anger and anxiety.

The personality of a person is described as a vector of five numeric scores, one for each trait, that traditionally have been calculated based on questionnaires filled by the person (McCrae and John 1992) and, more recently, by computerized tools based on text analysis techniques (Tausczik and Pennebaker 2010).

Linguistic Inquiry and Word Count, or simply **LIWC** (pronounced "Luke"), is the most prominent tool that derives psychological dimensions from the words of a text (Pennebaker et al. 2015). LIWC consists of a dictionary and a text processing module.

In the latest version (2015), the dictionary consists of up to 6400 tokens (i.e., words, word stems and selected emoticons) organized into 93 hierarchical *categories*, each collecting tokens that serve the same function. For instance, the category *Impersonal pronouns*

includes such tokens as "it", "it's" and "those". Importantly, a token may be contained in more than one category; for instance, the word "cried" is associated with five categories, namely *sadness*, *negative emotion*, *overall affect*, *verbs*, and *past focus*. The categories are classified along four *dimensions*: summary language variables (e.g., categories *Clout* and *Analytical thinking*), linguistic dimensions (e.g., categories *Impersonal pronouns*, *Auxiliary verbs*), other grammar (e.g., categories *Common verbs*, *Interrogatives*) and psychological processes (e.g., categories *Positive emotion*, *Negative emotion*).

For each word in a text, the text processing module increments the count of the categories that are associated with that word. At the end of the process, the text is represented as a vector of 93 scores, one for each category.

Receptiviti is the commercial face of LIWC that provides a powerful RESTful API, referred to as the Receptiviti People API, to obtain the psychometrics of the words of a text. Most interestingly, the API does not output only the values of the LIWC categories, but also the values of additional psychological measures, referred to as *Receptiviti measures*, that give precious insights into the psychological processes hidden in the words.

These Receptiviti measures are grouped into five psychological insights (Table 3) and are output by the API as raw scores on a scale from 0 to 100 (or, alternatively, on a 5-point scale). Each of the five measures of the Big Five insights also consist of 6 additional measures that are meant to give a fine-grained detail of the respective personality traits (Table 4). Additionally, Receptiviti provides a set of measures that do not belong to any aforementioned psychological insight and are classified as *Interests and Orientations* (Table 5); these measures quantify the orientations of a person towards things (e.g., money, food) and aspects of their lives (e.g., religion, work).

As of the time of writing, Receptiviti can analyze texts in English and Spanish, but the scores of the Receptiviti measures are only output for English texts, which is the reason why in the remainder of this section we use our dataset MonoDS.

5.2 Data exploration

In order to corroborate the hypothesis that the personality traits are good predictors of a user's interests, we used the Receptiviti People API to obtain three feature vector representations of the users in our own dataset MonoDS. First, we preprocess the text of the tweets of all users as explained in Sect. 4.2.1, that is we remove all stopwords, numbers, mentions and hashtags. Then, the tweets of each user u are passed to the Receptiviti API, so as to have a vector representation of u in terms of the LIWC and Receptiviti dimensions.

In the first representation, that we call Big5, a user u is represented by a vector of five dimensions, one for each Receptiviti measure in the Big Five insight described in Table 3. The idea here is to evaluate the predictive power of the five traits of the Big Five model alone. In the second representation, that we call Big5 $^+$, each user is represented by a vector of 35 dimensions, the five dimensions of the Big5 representation and the 30 measures associated to the Big Five traits described in Table 4. The idea is to measure the impact of the additional measures on the predictive power of the five traits. Finally, in the third representation, called Recep, the feature vector representing a user consists of all Receptiviti measures.

In order to summarize and visualize the important information in our dataset, we apply a technique called principal component analysis (PCA) to the three representations. For each representation, the input to PCA is the set of all vectors that represent all users in the

Table 3 Receptiviti measures

Insight	Measure	Definition
Cognitive thinking style	Thinking style	Degree to which the person is an analytical thinker
	Persuasive	Degree to which a person is able to persuade others
	Reward bias	Degree to which a person weighs risks vs. rewards when making decisions
Big five	Openness	Degree to which a person is open to new ideas and experiences
	Conscientiousness	Degree to which a person is reliable
	Extraversion	Degree to which a person feels energized and uplifted when interacting with others
	Agreeableness	Degree to which a person is inclined to please others
	Neuroticism	Degree to which a person expresses strong negative emotions
Social style	Social skills	Degree to which a person feels at ease with others
	Insecure	Degree to which a person lacks confidence when dealing with others
	Cold	Degree to which a person is emotionally unresponsive
	Family orientation	Degree to which a person values are rooted in their sense of family
Emotional style	Adjustment	Degree to which a person is able to maintain quality relationships with others.
	Happiness	Degree to which a person is optimistic
	Depression	Degree to which a person may have difficulty finding joy in their life
Working style	Independent	Degree to which a person is not a conformist
	Power driven	Degree to which a person is driven by the desire of power
	Type-A	Degree to which a person is competitive
	Workhorse	Degree to which a person has a strong work ethic

Excerpt of a table from *Receptiviti API—User Manual* that can be obtained at https://www.receptiviti.ai/ upon registration

Table 4 The additional measures of the five personality traits

Trait	Measures
Openness	Artistic, intellectual, liberal, imaginative, emotionally aware, adventurous
Conscientiousness	Self-assured, disciplined, ambitious, dutiful, cautious, organized
Extraversion	Sociable, friendly, assertive, energetic, cheerful, active
Agreeableness	Generous, trusting, cooperative, empathetic, genuine, humble
Neuroticism	Impulsive, stressed, anxious, aggressive, melancholy, self-conscious

Table 5 The additional measures for interests and orientations

Measure	Definition
Friendship focused	Degree to which a person focuses on friendship
Body focus	Degree to which a person focuses attention on their or others people's body
Health oriented	Degree to which a person is focused on health
Sexual focus	Degree to which a person focuses on sexuality
Food focus	Degree to which a person focuses on eating and drinking
Leisure oriented	Degree to which a person thinks about leisure activities
Money oriented	Degree to which a person thinks about money and finances
Religion oriented	Degree to which a person thinks about religion
Work oriented	Degree to which a person is focused on work or school
Netspeak	Degree to which a person is comfortable communicating with internet slang

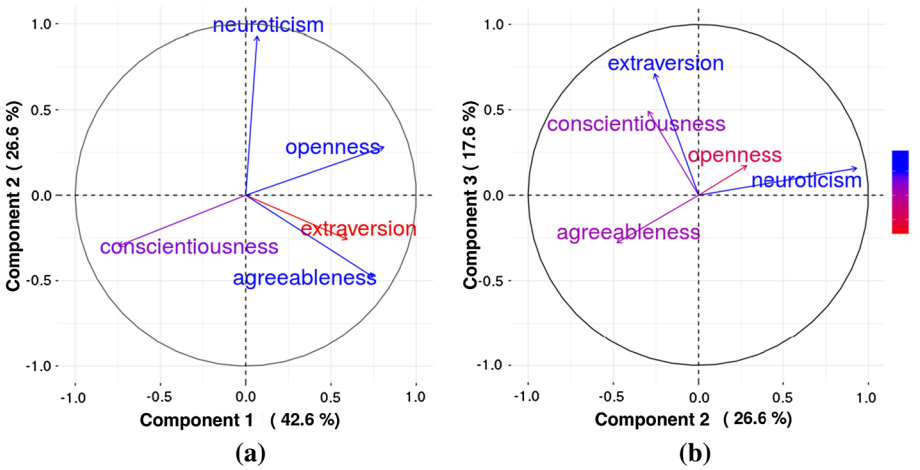


Fig. 2 PCA analysis of the representation Big5

dataset and the output is the same representation in a lower-dimensional space, where the new dimensions—called *principal components*—are a linear combination of the original.

This analysis is particularly helpful to unveil and visualize (Fig. 2) the relatedness of each dimension to the others in the original representation.

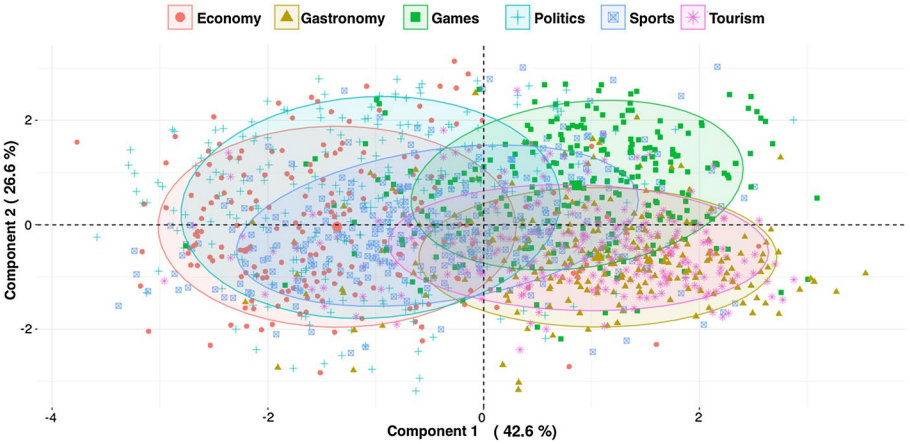


Fig. 3 Plot of the users on the first two components of BIG5

Figure 2a represents the two first components—component 1 on the x-axis and component 2 on the y-axis—obtained on BIG5. The percentages close to each component indicate the amount of information carried by each component; one can easily see that the first three components (see Fig. 2b, where the component 3 is represented along the y-axis) carry 86.8% of the information of the original representation and, therefore, they are expressive enough to describe the original representation. Each vector in the figure is a feature of BIG5 (e.g., openness, extraversion) and its color indicates whether the feature is well-described (color blue) or not (color red) by the two components. Thus, the features openness and agreeableness are well-described by the first two components (they are in blue in Fig. 2a) extraversion is not well described by the first two components (it is in red in Fig. 2a), but it is well described by the components 2 and 3 (it is in blue in Fig. 2b). The two figures also show that some features point to the same direction and form a small angle, which means that they are somehow related, while others point to opposite direction and form large angles, in which case they are weakly or not related. In particular, openness and agreeableness seem to be related because they both point to the positive quadrant of the first component in Fig. 2a. This is logic, as we would expect that a person open to new ideas and experiences is also someone agreeable. Note that the same two features are quite far apart in Fig. 2b, but nothing can be concluded on this because both features are not well described by the components 2 and 3; in other words, the relatedness between two features with respect to two components can be measured only if the features are both well described (i.e., they are in blue) by the components. Interestingly, neuroticism is not related to agreeableness (Fig. 2a) nor to extraversion (Fig. 2b), which is again to be expected.

If we plot the users of MONODS classified by their interest along the same components, we can study the correlation of the interests with the personality traits. Figure 3 shows the users plotted against the first two components. Some interesting observations can be made here. The vast majority of the users interested in the gastronomy and tourism concentrate on the fourth quadrant of the Cartesian space, where the feature agreeableness is oriented (Fig. 2a); a good portion of them are also found in the lower part of the first quadrant, where the feature openness is oriented. Only few outliers are in the higher part of the first quadrant, where the feature neuroticism is oriented. We can conclude that the users interested in gastronomy and tourism have a clear tendency to being open and agreeable

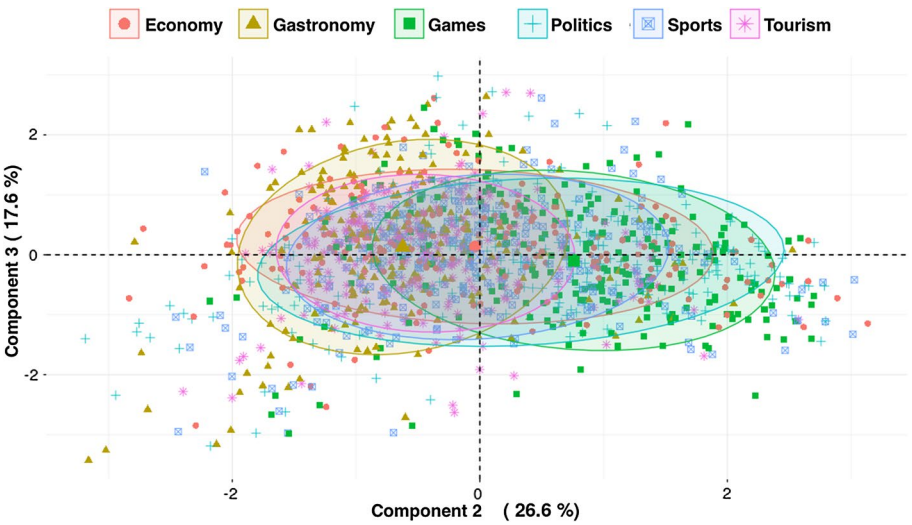


Fig. 4 Plot of the users against component 2 and 3 of BIG5

and less prone to negative feelings, which is somehow understandable because, given their interests, they might fill their lives with social activities involving trips and good food.

Interestingly, the two categories of users interested in politics and economy occupy mostly the second and third quadrants, opposite to the users interested in gastronomy and tourism. Here the tendency is more towards conscientiousness, which again is understandable. Some people follow the political and economic situation of their countries in order to make better decisions for themselves and their families. Most noticeably, many of such users display a mild tendency to neuroticism, in the same way as users interested in video games who, however, seem to be more open than conscientious. Finally, the users interested in sports lean towards openness and conscientiousness.

Looking at the Fig. 4, that shows the users plotted against the components 2 and 3, we can confirm these observations. Indeed, the users interested in politics and economy are oriented in the same way as the vector neuroticism in Fig. 2b; the users interested in gastronomy and tourism are between the vectors agreeableness and conscientiousness, although the former have sensibly higher values for the component 3, which suggests a tendency towards extraversion.

Similar considerations can be applied by looking at the graphs obtained by using the representations BIG5 $^+$ and RECEP. One important caveat is that the analyses presented here do not control for age and gender, which are not information that are easily available from Twitter.

5.3 Interest determination

We formulate the determination of the interests of a user u as a multinomial classification problem; each user is described as a feature vector, each feature being a Receptiviti measure, based on either of the three representations BIG5, BIG5 $^+$ or RECEP and the class being one of the interests in $\mathcal{I}$. We tried three different prediction models, multinomial logistic regression with ridge regression, SVM and Naive Bayes. Multinomial logistic regression

is a generalization of logistic regression to a multinomial classification problem where the classes are categorical, as in our case. Through multiple experiments, we found out that the SVM had a better prediction accuracy with a RBF kernel, with parameters $C = 10$ and $\gamma = 1.0$, determined through grid search.

6 Evaluation

6.1 Evaluation metrics

We use precision (P_i), recall (R_i) and f-measure (F_i), as the standard metrics to evaluate the prediction ability of an algorithm, for any given interest $i \in \mathcal{I}$:

$$P_i = \frac{|TP_i|}{|TP_i| + |FP_i|} \quad R_i = \frac{|TP_i|}{|TP_i| + |FN_i|} \quad F_i = \frac{2 \times P_i \times R_i}{P_i + R_i}$$

where:

- TP_i (true positives) is the set of users whose actual and predicted interest is i;
- FP_i (false positives) is the set of users whose predicted and actual interest are i and j respectively, $i \neq j$;
- FN_i (false negatives) is the set of users whose predicted and actual interest are j and i respectively, $j \neq i$.

In words, the precision measures how many of the predictions for the interest i are correct; the recall measures how many users whose actual interest is i are correctly predicted; the f-measure is a trade-off between precision and recall and an indication of the overall accuracy. The overall precision P, recall R and f-measure F of an algorithm is obtained by averaging P_i, R_i and F_i respectively over all interests in $\mathcal{I}$.

6.2 Evaluation of Frisk

As discussed in Sect. 4.2.2, we have two versions of Frisk, namely FriskLA, that uses a language-agnostic heuristic to obtain the bag-of-articles from the tweets of a user, and FriskTM, that uses TagMe. FriskLA is the multilingual version of Frisk, in the sense that it can discover the interests of a user without the need of specifying the language as an input. On the other side, FriskTM needs the language as an input. As a result, FriskLA is the only approach that we can evaluate on the multilingual dataset MultiDS.

Evaluation of FriskLA *on* MultiDS. The prediction ability of FriskLA depends on the number of tweets that are necessary to determine the interests of a user. Intuitively, the higher the number of tweets, the better the prediction ability. This is confirmed by our experiments as both the overall precision P and recall R increase when the number of tweets analyzed by FriskLA increases (Fig. 5), given that more words, and therefore more Wikipedia articles, are obtained from them. Remarkably, precision and recall have relatively high values (respectively, 0.84 and 0.80) even when a small amount of tweets (50) is considered, which indicates that FriskLA is particularly effective in determining the interests of users who do not tweet very often. Figure 5 also shows that considering more than 250 tweets will not affect the precision (which is stable after that point) while

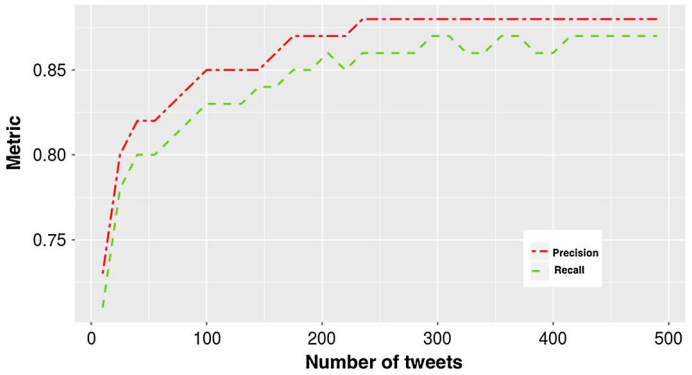
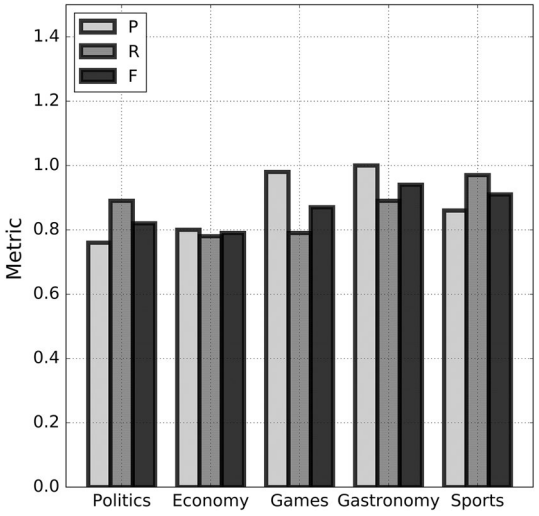


Fig. 5 Variation of precision and recall with the number of tweets

Fig. 6 Results of FriskLA on MultiDS



it contributes to a small increase of the recall. Based on these observations, we set the number of tweets to 350.

FriskLA obtains good values of precision and recall across all interests (Fig. 6). The two interests where the f-measure is lower are politics and economy. A closer look at the confusion matrix (not displayed here), reveals that FriskLA determines 322 users interested in politics, 77 of which are false positives (i.e., users with interests other than politics); most of these false positives (59) are users interested in the economy. This is probably due to the fact that politics and economy are highly related and it is not illogic to think that a user interested in politics is interested in the economy as well. As a matter of fact, the analyses of Figs. 3 and 4 show that users interested in politics and those interested in economy are quite similar in terms of personality traits. Sports and games are somehow related and, indeed, most of the false positives, when considering the users interested in sports (resp., games), consist of users interested in games (resp., sports); gastronomy is unrelated to the other interests, which explains why the high value of precision. In summary, these results indicate a good ability of FriskLA of

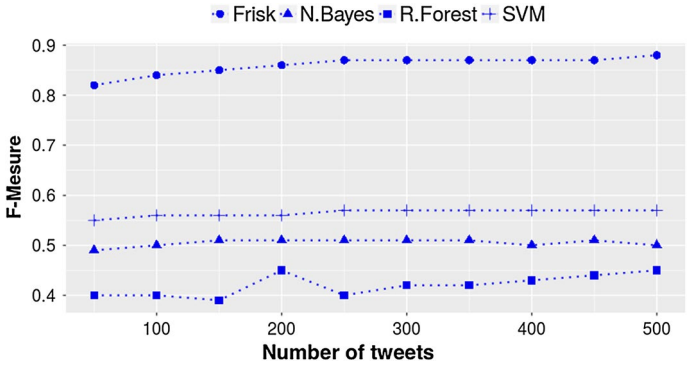


Fig. 7 F-measure of FriskLA and the text classifiers on MultiDS

predicting the interests of the users, even when the interests are highly related, as in the case of politics and economy.

Comparison against text classifiers. In a recent work, Raghuram and colleagues proposed a supervised approach to categorize the interests of Twitter users by using traditional classifiers (Raghuram et al. 2016). Here we want to verify whether the promising results that they reported in a monolingual context (which are coherent with our findings described below) can be generalized to a multilingual context. In order to train the text classifiers, we randomly split MultiDS to obtain a training set with 753 users and a test set with 594 user. In either set, an instance is a Twitter user represented as a tf-idf vector obtained from that user tweets, after removing stop words and special characters as explained in Sect. 4.2.1. Each user is assigned to one class, representing the interest of the user. In both sets, users are uniformly distributed over the five classes (to avoid class imbalance) and across the four languages.

We used scikit-learn, a popular machine learning library in Python, to train several text classifiers, of which we selected the best three: a Support Vector Machine (SVM) with linear kernel and hyper-parameter $C = 1.0$; a Multinomial Naive Bayes classifier with Laplacian smoothing parameter $\alpha = 1.0$; a Random Forest using the Gini impurity as the function to measure the quality of a split. Understandably, the effectiveness of the text classifiers is impacted by the number of tweets used to train them; Fig. 7 shows that the F-measure of the three classifiers is below 0.6 even when using 500 tweets and we observed that any further increase does not lead to any improvement. On the other hand, FriskTM obtains a f-measure closed to 0.9 when using 400 and 500 tweets.

Comparison againt LDA. We compare FriskLA against the approach proposed by Weng and colleagues that apply LDA to discover the interests of Twitter users (Weng et al. 2010). To this extent, we used the collapsed variational Bayes approximation to the LDA objective (Asuncion et al. 2009) implemented in the Stanford Topic Modeling Toolbox[2] to learn a topic model with five topics (one for each interest) over the training set of MultiDS obtained previously for the comparison with the text classifiers. The learning phase is necessary to determine the best values of the LDA parameters, namely the Dirichlet prior on the per-document topic distributions, and the Dirichlet prior on

[2] https://nlp.stanford.edu/software/tmt/tmt-0.4/.

Table 6 Interests as determined with LDA

Interest 1	Interest 2	Interest 3
Jeu, bien	Day, finance	Politica, calcio
Faire, j'ai	Win, watch	Solo, sapevatelo
Merci, nouveau	Support, minister	Ecco, dopo
Monde, faut	Team, week	Prima, renzi
Après, contre	Release, world	Cosa, italia
Entreprises, français	Work, donald	Ora, anni
Soir, bonjour	Check, president	Roma, grazie
Jour, jeux	Big, campaign	Sempre, lavoro
ans, demain	clinton, top	Ancora, tramite
Interest 4	Interest 5	
Juego, nuevo	Videogames, food	
Economia, gracias	Xbox, trailer	
Millones, partido	Gameplay, nintendo	
Ver, ser	Playstation, foodie	
Día, españa	Foodporn, wars	
Juegos, gran	Super, play	
Mundo, mejor	Day, sale	
Ahora, gobierno	Lego, week	
Nueva, semana	World, gamedev	

the per-topic word distribution (for both, we found that the best value was 0.01). For all tweets, we lowercased the words and we removed the stop words in the four languages so as to avoid to include words that are used too frequently in the topic model. The learned topic model consists of five groups of words (Table 6), one for each interest; the groups are unlabeled, meaning that one has to guess which group corresponds to which interest based on the words that occur in it.

From the table one can see that some interests are easily recognizable based on the words that describe them; this is the case of "Interest 2" and "Interest 5" that correspond to *Politics* (cf. words such as "minister", "donald", "president" and "clinton") and *Games* (cf. words such as "videogrames", "xbox", and "nintendo") respectively,

However, the other groups of words do not seem to identify unambiguously the other interests. For instance, "Interest 3" includes words in Italian, some of which (e.g., "calcio" that means "soccer") indicate the interest *Sports*, others (e.g., "politica", that means "politics", and "renzi", the family name of the former Italian prime minister) indicate the interest *Politics*. "Interest 1" is a collection of words in French, some of them (e.g., "jeu" and "jeux" that translate to "game" and "games" respectively) indicate either the interest *Games* or *Sports*, others (e.g., "entreprises" that translates to "companies") indicate the interest *Economy*. Finally, "Interest 4" is a group of words in Spanish that again do not point unambiguously to any of the five interests that we selected.

We played with the parameters of LDA, especially with the number of topics, to try to obtain groups of words that would identify our five interests. In particular, based on the observation that each group of words in Table 6 is only in one language, we set the

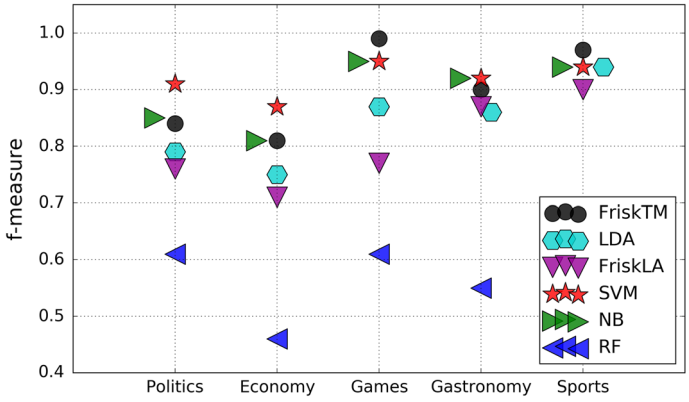


Fig. 8 Evaluation on MonoEN

number of topics to 20, in the hope to obtain four groups of words (one in each language) for each interest; however, we did not obtain that result.

In the end, we applied the learned topic model to the Twitter users in the test set of MultiDS and we computed the precision/recall/f-measure for the two interests *Games* and *Politics* that are identified unambiguously in the training phase. For the interest *Politics*, we obtain a precision/recall/f-measure of 0.29, 0.37 and 0.33 respectively; for the interest *Games*, we obtain 0.55, 0.34 and 0.42. In both cases, the values are well below those obtained with Frisk.

Evaluation on a monolingual dataset. We finally compare Frisk, the three text classifiers and LDA in a monolingual context, where the language of the users is known.

For this purpose, we split MultiDS into four datasets (MonoEN, MonoFR, MonoIT, MonoES), one per language, and each dataset is split in half, a training (used to train the three classifiers and a topic model) and a test set.

Figure 8 shows the f-measure obtained with the different approaches on the test set of MonoEN, that only includes users who tweet in English. The results on the other languages are similar and are omitted because of space constraints. The f-measure of FriskTM (represented as a black circle) is comparable to the values obtained in a multilingual context (with the only exception of *Games*), which indicates that Frisk does not depend on the language. Among the text classifiers, Random Forest (RF) performs poorly while SVM and Naive Bayes outperform Frisk across the five interests. The results seem to confirm the findings of Raghuram et al. (2016) who reported a good accuracy of traditional classifiers in a monolingual context.

Interestingly, the accuracy of FriskTMis comparable to that of SVM and Naive Bayes, which is probably the effect of TagMe considering the context of a tweet to assign an interpretation to its words. The only problem is that TagMe is not truly multilingual in the sense that it needs to know the language of the tweets in order to create $\mathcal{BOA}_u$.

Finally, LDA leads to a good accuracy across the five interests, comparable to or slightly better than, that of FriskLA but in general worse than SVM, Naive Bayes and FriskTM. We observe that, unlike the multilingual context, the topics discovered by LDA in a monolingual context are groups of words that unambiguously identify the five interests that we selected for the evaluation.

Table 7 Evaluation of ASCERTAIN on BIG5

	Politics	Econ.	Games	Gastr.	Sports	Tour.	Avg
Logit							
P	0.48	0.44	**0.58**	0.50	0.44	0.49	0.49
R	0.40	0.47	**0.74**	**0.48**	0.44	0.44	0.49
F	0.44	0.45	**0.65**	0.49	0.44	0.46	0.49
SVM							
P	**0.56**	**0.46**	**0.58**	**0.52**	**0.51**	**0.56**	**0.54**
R	**0.44**	**0.52**	0.71	**0.48**	**0.56**	**0.48**	**0.53**
F	**0.50**	**0.49**	0.64	**0.50**	**0.54**	**0.52**	**0.53**
Bayes							
P	0.49	0.37	0.53	0.46	0.43	0.45	0.46
R	0.32	0.51	0.63	0.40	0.39	**0.48**	0.45
F	0.39	0.43	0.58	0.42	0.41	0.47	0.45

Best values are highlighted in bold

6.3 Evaluation of ascertain

We evaluated ASCERTAIN on three representations, namely BIG5, BIG5 $^+$, and RECEP, as discussed in Sect. 5.2. For each representation, we experimented with three prediction models, multinomial logistic regression, SVM and Naive Bayes. We used the dataset MONODSand we obtained the results by using a 10-fold cross-validation.

The results obtained by using the representation BIG5 (Table 7) show that the prediction accuracy is quite low for the three classifiers, with SVM being slightly better. This suggests that while the five traits of personality have some degree of correlation with the interests, as shown in Figs. 3 and 4, these correlations are not To interpret these results, it is fair to remind that the five Receptiviti measures that correspond to the five traits of personality are a summary of fine-grained measures, six for each trait. Intuitively, five summary measures are too generic to have a significant predictive power. This is why we considered the representation BIG5 $^+$, which includes the five measures of BIG5 and also the additional 30 fine-grained measures detailing the five traits.

As we would expect (Table 8), the use of the additional features increases significantly the prediction ability of the three classifiers, the SVM still being the best one, although the logistic regression achieves a better precision for the interests economy, games and sports. Naive Bayes has a considerably lower accuracy across all interests. It is interesting to note that the three classifiers have the highest precision/recall and f-measure for the interest sports, indicating that the features selected in BIG5 $^+$are highly discriminant for this interest. On the other side, the lowest accuracy (in terms of precision, recall and f-measure) is obtained by the three classifiers on the interest gastronomy.

Taking a closer look at the confusion matrix (not displayed here), we observe that all three classifiers consistently tend to misclassify the users interested in politics as users interested in economy (and the other way round) and the users interested in gastronomy as users interested in tourism. Taking the example of logistic regression, 52% (resp., 35%) of the users misclassified as interested in economy (resp., politics) are interested in politics (resp., economy); 40% (resp., 38%) of the users misclassified as interested in gastronomy (resp., tourism) are interested in tourism (resp., gastronomy). Same considerations apply to the SVM and Bayes. This confirms what we already observed while evaluating FRISK on

Table 8 Evaluation of ASCERTAIN on BIG5 $^{+}$

	Politics	Econ.	Games	Gastr.	Sports	Tour.	Avg
Logit							
P	0.70	**0.75**	**0.77**	0.61	**0.79**	0.65	0.71
R	**0.70**	0.79	**0.75**	0.61	0.77	0.66	0.71
F	0.70	**0.77**	**0.76**	0.61	0.78	0.65	0.71
SVM							
P	**0.74**	0.73	0.69	**0.64**	0.78	**0.71**	**0.72**
R	**0.70**	**0.81**	0.66	**0.62**	**0.82**	**0.69**	**0.72**
F	**0.72**	**0.77**	0.68	**0.63**	**0.80**	**0.70**	**0.72**
Bayes							
P	0.62	0.64	0.56	0.46	0.77	0.50	0.60
R	0.57	0.73	0.58	0.48	0.65	0.54	0.59
F	0.59	0.68	0.57	0.47	0.71	0.52	0.59

Best values are highlighted in bold

Table 9 Evaluation of ASCERTAIN on RECEP

	Politics	Econ.	Games	Gastr.	Sports	Tour.	Avg
Logit							
P	**0.77**	**0.76**	**0.74**	0.86	**0.84**	0.75	**0.79**
R	**0.74**	0.78	**0.75**	**0.88**	0.84	0.73	**0.79**
F	**0.76**	**0.77**	**0.74**	**0.87**	**0.84**	**0.74**	**0.79**
SVM							
P	0.73	0.73	0.68	**0.87**	0.81	**0.77**	0.77
R	0.69	0.78	0.72	0.86	**0.85**	0.69	0.77
F	0.71	0.76	0.70	0.86	0.83	0.73	0.77
Bayes							
P	0.70	0.69	0.59	0.86	0.82	0.54	0.70
R	0.59	**0.79**	0.62	0.65	0.67	**0.76**	0.68
F	0.64	0.74	0.61	0.74	0.74	0.63	0.68

Best values are highlighted in bold

the multilingual dataset, concerning the similarity of the interests politics and economy. Moreover, this also confirms the observations that we made in Figs. 3 and 4, concerning the fact that the users interested in politics display similar personality traits as the users interested in the economy and the users interested in tourism display similar personality traits as the users interested in gastronomy.

Finally, Table 9 show the results obtained by the three classifiers by using the representation RECEP that includes all the Receptiviti measures. In this case, the logistic regression is the best in terms of precision, recall and f-measure: more in details, the logistic regression achieves the best f-measure across all interests. In general, all three classifiers have a better prediction accuracy than when considering the first two representations. This means that the all the Receptiviti measures combined contribute to a better discrimination of the users based on their interests. Naive Bayes has the best recall for economy and tourism, but in general these results have to be considered as two outliers, because in the other interests the recall is always lower than 0.70; SVM and logistic regression are more stable across all

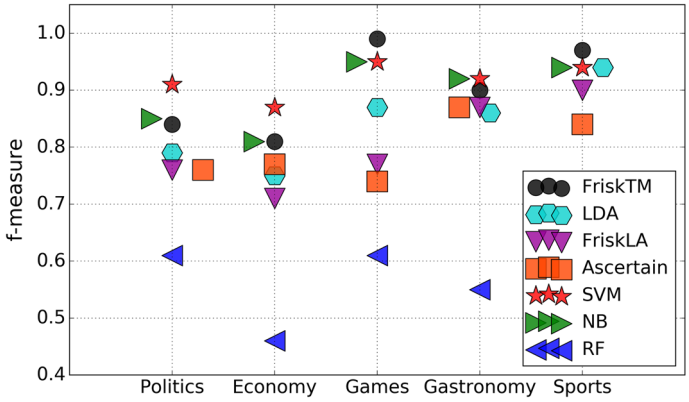


Fig. 9 Comparison of all approaches

interests. Importantly, we can make the same observations on the misclassification errors concerning the users interested in politics and economy and the users interested in gastronomy and tourism.

6.4 Comparison and discussion

In Fig. 9 we compare the results obtained by Ascertain, Frisk and all the other techniques that we have discussed on the dataset MonoEN. The figure shows the f-measure obtained with all the approaches for each interest; for each approach, we use the best configuration of parameters, which we recall here:

- Frisk: minimum number of categories $\theta = 5$ and maximum ancestor $k = 3$. We use the two variants of Frisk, namely FriskLA and FriskTM, the first using our custom disambiguation algorithm when creating the BOA and the second using TagMe.
- Ascertain: logistic regression classifier trained on Recep.
- LDA: five topics, Dirichlet prior on the per-document topic distributions, and Dirichlet prior on the per-topic word distribution set to 0.01
- SVM: linear kernel, hyper-parameter $C = 1.0$.
- Multinomial Naive Bayes. Laplacian smoothing parameter $\alpha = 1.0$.
- Random Forest. Gini impurity as the function to measure the quality of a split.

Interestingly, the values of the f-measure obtained by Ascertain follow the same pattern as the values obtained by the other approaches with higher values for the interests Gastronomy and Sports. Overall Ascertain is comparable to FriskLA, indeed they have the same average f-measure over the five interests, while FriskTM is consistently superior across all interests. We already pointed out that FriskTM is understandably better than FriskLAin a monolingual context, because it uses a better disambiguation mechanism, albeit not a language-agnostic one. Since Ascertain can only be applied to tweets written in English (Receptiviti only analyzes English texts), we will now base the discussion on the comparison between Frisk and Ascertain by using FriskTM as a reference.

One of the contributions of this paper is the comparison between an approach (Frisk) that infers the interests of social media users by using the explicit semantics of the words

and an approach (Ascertain) that uses the hidden psychological dimensions of the words. Understandably, the first is more effective because it establishes a direct link between the words used in the tweets and the topics of interest. After all, when one person uses several words such as *election*, *parliament*, *presidency* and *government*, s/he clearly reveals her interest in politics.

Considering that Ascertain does not use the semantics of the words at all, but only the hidden psychological dimensions as computed by Receptiviti, its accuracy is rather impressive. In other words, the way people express themselves does not only reveal insights into their personality, but also into their leisure and professional activities. This also implies that Ascertain would be able to infer the interests of a person even if that person does not mention them explicitly, something that Frisk, or any other approach that is based on the explicit semantics of the words, could not do.

7 Concluding remarks

In this paper, we presented two approaches to the automatic determination of interests of social media users.

The first approach, named Frisk, determines the interests of a user from the explicit meaning of the words that occur in the textual posts of the user. The second approach, named Ascertain, determines the interests by using the psychological dimensions hidden in the words, as computed by the tool Receptiviti. To the best of our knowledge, Frisk is the first unsupervised multilingual approach that gives a precise categorization of the interests, in the same way as supervised approaches do but without the need of a training set. Also, Ascertain is the first approach that exploits the psychological dimensions of the words to infer the interests of a person. The evaluation shows that both approaches have quantitative promise. Nevertheless, there are some areas of improvement.

First, Frisk is based on a simple language-agnostic heuristic to select the proper meaning of a word that occurs in a tweet. A more sophisticated language-agnostic disambiguation method is likely to result in a better prediction accuracy, as the evaluation in the monolingual context with TagMe suggests. Also, as of the current version, Frisk models the tweets as a bag-of-words and ignores the n-grams; the use of n-grams (e.g., "Attorney General") is key to capture precious information as to the interest of a person (e.g., politics). Although in this paper we focused on the prediction accuracy of Frisk, an evaluation of its performances is needed and we are currently studying a parallel implementation.

As for Ascertain, the methodology shows that the psychological dimensions as computed by Receptiviti are good predictors of the interests of a person. It would be interesting to verify if these findings generalize to other languages as well. As of the time of writing, the Receptiviti measures can only be computed for texts written in English. The correlations that we found need to be better assessed for a larger number of interests; also, a finer-grained study is necessary to control for age and sex, two factors that have usually a significant impact on the analysis of the correlations.

References

Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. (2009). On smoothing and inference for topic models. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence, UAI '09*, pp. 27–34. AUAI Press.

Bao, H., Li, Q., Liao, S. S., Song, S., & Gao, H. (2013). A new temporal and social PMF-based method to predict users' interests in micro-blogging. *Decision Support Systems*, *55*(3), 698–709.

Bhattacharya, P., Zafar, M. B., Ganguly, N., Ghosh, S., & Gummadi, K. P. (2014). Inferring user interests in the twitter social network. In *RecSys*, pp. 357–360.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Cantador, I., Fernández-Tobías, I., & Bellogín, A. (2013). Relating personality types with user preferences in multiple entertainment domains. In *CEUR workshop proceedings*. Shlomo Berkovsky.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, *41*(1), 417–440.

Ding, Y., & Jiang, J. (2014). Extracting interest tags from twitter user biographies. In *Information retrieval technology*, pp. 268–279.

Eysenck, H. J. (2012). *A model for personality*. London: Springer.

Ferragina, P., & Scaiella, U. (2010). Tagme: On-the-fly annotation of short text fragments (by Wikipedia entities). In *CIKM*, pp. 1625–1628.

Furnham, A., & Heaven, P. (1999). *Personality and social behaviour*. Arnold.

Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, *59*(6), 1216.

Gottfredson, G. D., Jones, E. M., & Holland, J. L. (1993). Personality and vocational interests: The relation of Holland's six interest dimensions to five robust dimensions of personality. *Journal of Counseling Psychology*, *40*(4), 518.

He, W., Liu, H., He, J., Tang, S., & Du, X. (2015). Extracting interest tags for non-famous users in social network. In *CIKM*, pp. 861–870. ACM.

Jipmo, C. N., Quercini, G., & Bennacer, N. (2017). Frisk: A multilingual approach to find twitteR InterestS via wiKipedia. to appear.

Kapanipathi, P., Jain, P., Venkatramani, C., & Sheth, A. P. (2014). User interests identification on twitter using a hierarchical knowledge base. In *The semantic web: Trends and challenges—11th international conference, ESWC 2014, Anissaras, Crete, Greece, May 25–29, 2014. Proceedings*, pp. 99–113.

Kazdin, A. E. (2000). Encyclopedia of psychology.

Li, X., Guo, L., & Zhao, Y. E. (2008). Tag-based social interest discovery. In *WWW*, pp. 675–684.

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

Michelson, M., & Macskassy, S. A. (2010). Discovering users' topics of interest on twitter: A first look. In *4th Workshop on analytics for noisy unstructured text data*, pp. 73–80. ACM.

Moro, A., Raganato, A., & Navigli, R. (2014). Entity linking meets word sense disambiguation: A unified approach. *TACL*, *2*, 231–244.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, *41*(2), 10.

Odic, A., Tkalcic, M., Tasic, J. F., & Kosir, A. (2013). Personality and social context: Impact on emotion induction from movies. In *UMAP workshops*.

Pennacchiotti, M., Silvestri, F., Vahabi, H., & Venturini, R. (2012). Making your interests follow you on twitter. In *CIKM*, pp. 165–174.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of liwc2015*. Tech. rep., Austin, TX: University of Texas at Austin.

Perrin, A. (2015). *Social media usage*. Pew Research Center.

Raghuram, M., Akshay, K., & Chandrasekaran, K. (2016). Efficient user profiling in twitter social network using traditional classifiers. In *Intelligent systems technologies and applications*, pp. 399–411.

Rawlings, D., & Ciancarelli, V. (1997). Music preference and the five-factor model of the neo personality inventory. *Psychology of Music*, *25*(2), 120–132.

Rentfrow, P. J., Goldberg, L. R., & Zilca, R. (2011). Listening, watching, and reading: The structure and correlates of entertainment preferences. *Journal of Personality*, *79*(2), 223–258.

Rentfrow, P. J., & Gosling, S. D. (2003). The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, *84*(6), 1236.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS One*, *8*(9), e73,791.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS One*, *8*(9), e73791.

Spasojevic, N., Yan, J., Rao, A., & Bhattacharyya, P. (2014). LASTA: Large scale topic assignment on multiple social networks. In *KDD*, pp. 1809–1818.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval, SIGIR '10*, pp. 841–842. ACM.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.

Thurstone, L. L. (1934). The vectors of mind. *Psychological Review*, *41*(1), 1.

Vu, T., & Perez, V. (2013). Interest mining from user tweets. In *CIKM*, pp. 1869–1872.

Wang, T., Liu, H., He, J., & Du, X. (2013). Mining user interests from information sharing behaviors in social media. In *Advances in knowledge discovery and data mining*, pp. 85–98.

Wang, X., Liu, H., & Fan, W. (2011). Connecting users with similar interests via tag network inference. In *CIKM*, pp. 1019–1024. ACM.

Wen, Z., & Lin, C. Y. (2011). Improving user interest inference from social neighbors. In *CIKM*, pp. 1001–1006.

Weng, J., Lim, E. P., Jiang, J., & He, Q. (2010). TwitterRank: Finding topic-sensitive influential twitterers. In *WSDM*, pp. 261–270.

Witten, I. H., & Milne, D. N. (2008). An effective, low-cost measure of semantic relatedness obtained from wikipedia links.

Xu, Z., Lu, R., Xiang, L., & Yang, Q. (2011). Discovering user interest on twitter with a modified author-topic model. *WI-IAT*, *1*, 422–429.

Zarrinkalam, F., Fani, H., Bagheri, E., Kahani, M., & Du, W. (2015). Semantics-enabled user interest detection from twitter. *WI-IAT*, *1*, 469–476.