

POLITECNICO DI TORINO

Master degree course in Computer Engineering

Master Degree Thesis

Micro Influencer Detector

From Marketing Parameters to Semantic Analysis



Supervisor

prof. Maurizio Morisio

Candidates

Simone LEONARDI

Student ID: 235086

Internship Tutor

LINKS

dott. ing. phd. Giuseppe Rizzo

Academic Year 2018 - 2019

This work is subject to the Licence as Described on Politecnico di Torino website

Contents

| | | |
|-----------|--|-----------|
| I | Abstract | 5 |
| 1 | The intuition behind this work | 7 |
| 2 | Early detection of micro influencer | 8 |
| 2.1 | New jobs creation | 8 |
| 2.2 | Title explained | 8 |
| 2.3 | Who is a micro influencer | 9 |
| 2.4 | How to find them | 10 |
| 2.5 | Results | 10 |
| 2.6 | Conclusion | 10 |
| 3 | Introduction | 11 |
| 3.1 | Awareness | 11 |
| 3.2 | Why this tool now? | 11 |
| 3.3 | What research questions we answered | 11 |
| 3.4 | How to read this document | 12 |
| 4 | Use Case Scenarios | 13 |
| 4.1 | Company side | 13 |
| 4.2 | Potential Micro Influencer Side | 13 |
| II | State of the art | 14 |
| 5 | Psychological studies on personality and social interactions | 16 |
| 5.1 | Sociology and Psychology | 17 |
| 5.2 | Schwartz' Basic Human Values | 17 |
| 5.3 | Five Factor Model | 19 |
| 6 | How text can be a valuable source for measuring personality traits and social indicators | 21 |
| 6.1 | Personality from survey | 21 |

| | | |
|------------|--|-----------|
| 6.2 | Personality from social media | 23 |
| 6.3 | How to avoid Panopticon? | 25 |
| 7 | Word embeddings | 27 |
| 7.1 | Comparison on techniques | 28 |
| 8 | How information propagates itself in a community | 29 |
| 8.1 | Groups Behaviour | 30 |
| 8.2 | Networks, Crowds, Markets | 31 |
| 9 | Other relevant studies on this field and obtained results | 35 |
| 9.1 | Studies on influencers figure | 35 |
| 9.2 | YouTube experiment at IDIAP | 36 |
| III | Early Detection of Micro Influencers | 37 |
| 10 | Mathematical derivations | 38 |
| 11 | Implementation details | 41 |
| 11.1 | Technical choices | 41 |
| 11.2 | Packages and libraries | 45 |
| 11.3 | Tweets Preprocessing | 45 |
| 11.4 | Learning model choice | 46 |
| 11.5 | Why we choose SVM | 46 |
| 12 | Experimental Setup | 48 |
| 12.1 | Dataset: Statistics and Insights | 48 |
| 12.2 | Validation metrics | 48 |
| 12.3 | Consideration on scoring results | 52 |
| 12.4 | Results | 53 |
| 12.5 | Discussions | 54 |
| IV | Conclusion and Future Work | 56 |
| | Bibliography | 59 |

Part I
Abstract

We perform an analysis on Twitter social network users, following economics parametrization of special candidates considered micro influencers. Micro influencers are people with a community based on 1k to 20k followers, which have high reputation in a certain topic, a sort of guru, having an extremely high engagement power on other users in the community and who, with their posts on social network, can push their followers to action. From these understandings, we retrieve tweets and other parameters of users considered potential micro influencers, we analyze their marketing parameters and then performe semantic analysis to derive their Big5 and Basic Human Values from tweets through machine learning approach. We perform a social and psychological investigation over these areas. We understand how information is propagated in a community and by which agents. We inspect the cascading information propagation behavior. We determine if a user, examined as described before, can be defined as micro influencers and then we validate our hypothesis through cross validation and use of Support Vector Machine and Convolutional Neural Network over the psychological traits excluding marketing parameters. The research is effective, and we demonstrate how human values and personality traits are correlated with the influencing effect. On one side, we have created a tool for companies to find and contact potential evangelists of their message and brand, and on the other side new job opportunities for social network users that can now monetize their passions.

Chapter 1

The intuition behind this work

What does push human being to action? How does information propagates itself inside a small community? We are living in an era overwhelmed by false prophets and fake news, but, at the same time, we struggle for meaningful advises. The same concept is worth for companies which are looking for evangelists or brand ambassadors with high ROI (Return On Investment), because if the concept of influencer is well known, the cost of their work maybe not as well. Nowadays there is an attempt to make marketing analysis fast converging to an output that reveals who is considered an influencer or not. In this work we develop a tool that goes beyond this previous goal and at the same time exploits sentiment analysis, psychological and sociological studies, linguistic approach, with the aim of retrieving micro influencers analyzing what a person writes. We prepare the basis for conversational agents able to discover with dialogues if a person is right for this job or not. We download tweets and followers list of users appearing in search query over Twitter platform, giving a specific hashtag (topic), filter out the one not having right marketing parameters. After this we compute three main scores, again related to marketing definition of micro influencer, and we give a binary label to each user. In parallel we compute Big5 and Basic Human Values from both Five Factor Model [3] and Schwartz studies[2]. We finally fit the model excluding marketing values and predict output using only Big5 and Basic Human values. The model can certainly be improved, but we place the basis for future development where micro influencers can be directly detected from their writings avoiding the social media platform limitations in term of time restrictions and at the same time assuring final user of what data we are really and exactly using.

Chapter 2

Early detection of micro influencer

2.1 New jobs creation

All of us saw and lived economical crisis of 2008, some of us are still living it now. But as said:

Creativity comes from anxiety as the day comes from the dark night. [...] In the crisis that is inventiveness, discoveries and great strategies. Who overcomes crisis overcomes himself without being 'passed' [A. EINSTEIN, The crisis according to Albert Einstein (1955)]

With this in mind, all our work is devoted to find new solution for old problems. Because if the questions can remain equal, the answers need to change. We aim to create new job opportunities from one side and new cheaper and effective way of performing marketing or sensitization for companies on the other. With our tool, companies can early discover micro influencers (later detailed), that with respect to macro influencer or legends with more than 500k followers have much lower engagement cost[30], in large enough number creating sustainable and both side profitable job opportunities.

2.2 Title explained

Early detection of micro influencer means to find someone who has the potential of becoming the figure neede but neither she doesn't know yet neither

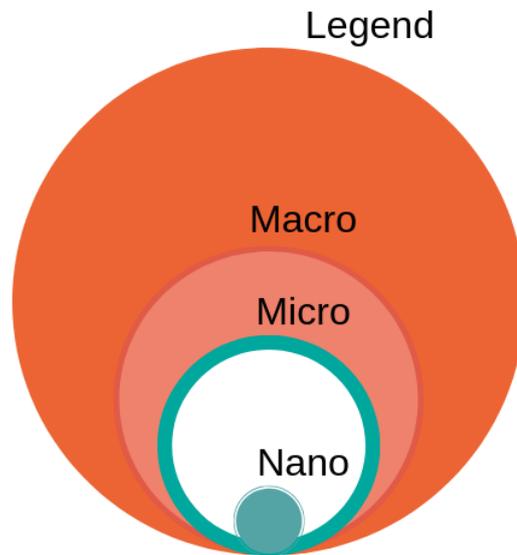


Figure 2.1: Influencer categories number of followers based

she is engaged from a company yet. This is profitable both for users and companies because we propose a contact to the potential micro influencer even if he is not aware of this possibility of income based on his passions and abilities in writing and we suggest candidates to companies that can hire many micro influencers with lower costs but high possibilities of efficient marketing campaigns. If we highlight the word early, we will notice that we enter in a trend at its beginning with all the chances of having high profitability in this situation.

2.3 Who is a micro influencer

According to Forbes definition¹ micro influecer are people with a smaller community of followers ranging in numbers from 10k and 500k, but we considered an average of multiple definitions found navigating the web and arriving at the final filter of 1k and 20k followers. But this reasoning still matter when we consider the other characteristics of this figure: high engagement with the community, high values in number of feedback received on his social media posts, overlapping of followers with other micro influencers talking on similar

¹<https://www.forbes.com/sites/barrettwissman/2018/03/02/micro-influencers-the-marketing-force-of-the-future/>

topics. The main differences between the already known standard influencer (i.e. Chiara Ferragni) are: number of followers, high reaction of followers in terms of action, low costs per post published, considered trusty in a certain topic because described as a passion and not just because paid to say things.

2.4 How to find them

We exploit the Twitter platform making query through REST API in Tweepy library for Python. We start from hashtag (topic) and later retrieve a list of user with enough follower and not spammer (people who follow other users just to be followed back) that posted on that topic, we later compute statistics on their platform numbers and decide which are potential candidates to the position of micro influencer. It would be interesting to explore other social media platforms such as Instagram, Facebook and YouTube. Twitter introduce the randomness in the process when we find users because we receive the ones that are talking about the specific topic recently, after search query, but having enough time we can make a storyboard to see how many times we retrieve same users. The time is monitored with our tool with the scores we compute but they are over whole corpus of tweets posted by users, instead future works can focus their investigation also on timestamps related to each tweet to see the consistency of a user over certain topics.

2.5 Results

We analyzes more than five million tweets and over one thousand user with plus than 10 topics and we develop a predictor model that must be improved because even if it has high recall and accuracy is still low in precision. In the results section of the thesis there is a comparison between learning models used.

2.6 Conclusion

We can conclude this synthesis saying that great step forward in the right directions has been made but even more need to be performed to improve the tool and also increase the common knowledge over the micro influencer figures and his related psychological parameters. At the same time highlighting the focus over the mechanism with which people can be sensitized and not manipulated. You can find the mirco influencer detector tool on Github ².

²github.com/D2KLab/micro-influencer-detector

Chapter 3

Introduction

3.1 Awareness

As suggested by David Foster Wallace, we must be aware of what we are living in and what are our purposes. The topic covered along this thesis work can have a huge impact on human life in terms of profiling issues and ethics, so we want the reader as the composer is aware that the intention is to give firstly to the end user a way to better understand itself and its potential and just later to the companies, with user consensus, to exploit the possibilities unchained by this paradigm. I'll talk about human personalities parametric and micro influencer discovery[22].

3.2 Why this tool now?

This tool is useful now both for the economical situation and for the possibilities it unlocks. We are at exact moment where the complexity of nowadays problems require the multidisciplinary included in this tool. Users tend to generate lot of noise in social media platform we do not use in target acquisition, and so it is mandatory to operate a search and a filter through this kind of smart automatizing. We learn from marketing and social behaviour what are the needs under the surface and we tackle the difficulties by creating new technologies.

3.3 What research questions we answered

First: Is it possible to define a micro influencer detector starting from a topic request from a company? Second: Is it possible to correlate word embeddings

and personality parameters from writing to the output of marketing score definition of micro influencer to avoid the step of user data retrieval but just reading her writings?

3.4 How to read this document

This document is divided in four main parts so if you are already aware of some of them you can skip directly to the topic you want to discover. The main topics are: Five Factor Model, Schwartz Basic Human Values, How information propagates itself in a social graph, mathematical derivation and implementation of our tool over Twitter platform.

Chapter 4

Use Case Scenarios

4.1 Company side

The marketing expert of a company approach our tool to satisfies his need of finding a brand ambassador but without spending lot of money just toward one big influencer. At the same time she looks for influencer passionate on her company treated topics so we can perform an investigation over some hashtag correlataed to the company message or product. If the client ask for topic we already searched we can perform a fast search on our database and in parallel perform a new search on that topic if outdated. We can tune filter followinf client requests both for follower and threshold. This approach remain marketing oriented. If the topic has never been serached we can perform a detection from scratch. In both cases we give to the marketing expert a list of potential micro influencers.

4.2 Potential Micro Influencer Side

A person who want to find a new source of income through his passions approach our tool asking to be analysed in his potetial over certain topics. We insert him in a list of users of specified topics and confront his values with the already searched one. Also in this case we can perform a new topic search from scratch, it will just need more time. At the moment we must pass through his platform numbers concerning followers and topic, but in a later stage of this tool we can just conversate with him over the specific topic and understand his personality traits and say if he is or can be a potential micro influencer.

Part II
State of the art

We start a path that converges two different approaches: machine learning based on numerical dataset and sentiment analysis that has as its tools writings retrieval and word embeddings on traces of semantic understanding. The intuition behind this new hybrid way of approaching the system is that data retrieval asking information to social media platform is both less ethic and not completely transparent and they are really hard to execute due to social media platform restrictions. It could be an example the attempt of retrieving large amount of data to develop a robust prediction model from Twitter with the limitation on Rate Limit Error that puts to sleep the REST API request for 15 minutes, involving a tremendous loss of time. We do understand the reasons behind these limitations (server occupation, avoidance of sensitive data leak, etc.) and so we want to make a model that can bypass these restrictions working on direct approach with user that want to demonstrate himself as a micro influencer "talking" with us and letting us analyze his writings for his personal opportunity of obtaining a job in marketing for topics he really enjoys. For what concern psychological data, instead, even if short questionnaire has leveraged the human inertia on performing boring action, there are many rejection on the objectivity of these auto assessment report[54]. In consequence many linguistics in collaboration with psychologists and sociologists developed the embedding model with semantic and syntactic clustering[25], [26], [27]. We take these work of correlations and labelling and training on data to make another step forward in analysis simplification: give me what you write and I tell you if you are a micro influencer or if you could begin one of them. First studies on detecting semantic and related psychological values score from writings date back to the middle of 19th century when Sir Francis Galton started this association[55]. He begun to create a dictionary (the ancestors of nowadays word embedding) of the English language clustering words with semantic similarities with in mind the goal of extracting significative association in words use in a subject writings. Later Allport and Odbert in 1936 better define groups of word separating them in *columns* and improving the semantic association with behaviour and psychology terms[16]. Just later Walter Norton following the path of his predecessors and merging the work with new discoreries, applied *Factor Analysis* to better understand semantic proximity of word clusters. And doing so he refined the model improving the definition. All this work later opened the doors to Five Factor Model on the work of Peabody and Goldberg[34].

Chapter 5

Psychological studies on personality and social interactions

Following Umberto Galimberti thoughts, Psychology is subjective by itself and every attempt of making it objective in the pursuit of creation of a psycho science will be deleterious. So it would be useful to define a new in between science or field that can make IT community, sociologists and psychologists working together, like in an Asimov's bestseller, we could name it "psycho historiography". This premise is useful when the community has to decide which is the way to follow. Till now many approaches are birth and growth [15]: many questionnaire and studies on personalities assessment: Five Factor Model, Schwartz on Basic Human Values, 16 personalities factor, across many countries and scientific areas[16], [17], [18], [2]. There are many main ways of retrieving personalities parameters psychologically speaking: Portrait Value Questionnaire with Likert Scale Questionnaire, Schwartz Value Survey [19]. There is NEO-PI-R from McRae and Costa [20] questionnaire. These methods are self reporting and subjected to partial observation of values from within, but this drawback are acceptable and embedded in the model. On the linguistic side instead we have embeddings and semantic cluster of words crated by sintactic and semantic affinity in context of writings [21]. Analysis of parameters and values are conducted on candidates' writings over as long as possible period of time, so evolving a chronicle of the subject, but at the same time his parameters are extremely influenced by contemporaneity and the moment of his life passing by. Considering this two aspects allows to calibrate the weight of them. We can make a parallelism with the Net based on historical data, but giving more weight to what has happened in instants immediately before the actual one. It is important to keep in mind that

Brian Little [53] says that typologies are the objective characteristics of an individual, but the most important for human being is the implicit or free aspects of personality, the ones that make ourselves unique, so the goal we have in lives. On the other side Big5, Basic Human Values and embeddings are useful for grouping similar tendencies in human behaviour and extract strategies to improve their quality of life.

5.1 Sociology and Psychology

This thesis would be a concretization of multiple subjects approach to solve a complex problem. In particular, we locate ourselves in the intersection of computer engineering, sociology and psychology. Concerning sociology we analyzed the interaction among social group, in particular in digital social networks environment, with the goal of understanding how information propagate itself over the social graph and so in the communities environment. On the other side, reasoning about psychology, we touched the Five Factor Model hot topic and we got in contact with the human values theory of Shalom Schwartz. In order to apply this two previous topic in a scientific methodological path we used deep learning approach, topic classification, statistical analyses, and machine learning algorithm from computer engineering field.

5.2 Schwartz' Basic Human Values

The theory developed throughout 90's from Shalom Schwartz aims to identify 19 (10 originally) human values spread and shared all over the world among different cultures. He investigated what behaviours are behind those emerged values and how some of them highlight particular aspects with respect to the others. Schwartz, in order to empirically supports his thesis, performed surveys [2] in 82 different countries and with distances of decades he updated and redid them again with upgraded parameters. Schwartz discovered that individuals and groups have different hierarchies of values, based on localization, political environment, human rights and so on. The other problem tackled was to use empirical and numerical, so reliable, measuring methods that, following the scientific one, can be repeated and compared among social scientist and the rest of the research world. We have to agree about what values are and why we exploit them in our thesis. Values are treated in Schwartz's analysis as believes and common goals that produce actions in members of the group sharing them. Values go further laws, they

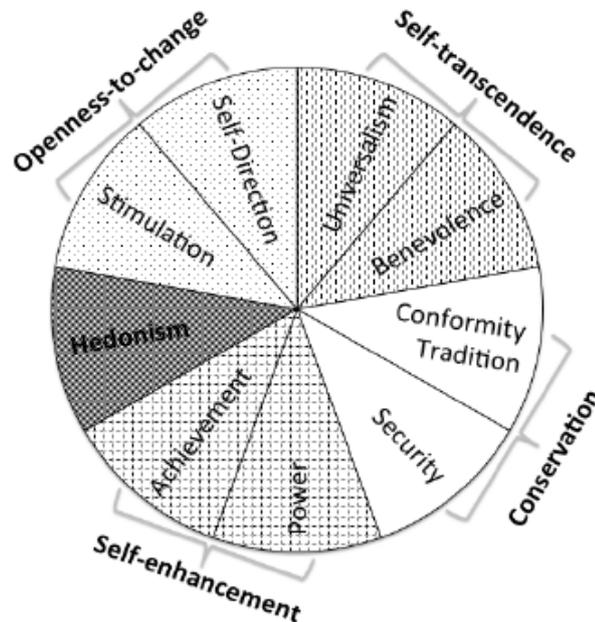


Figure 5.1: Schwartz' Values

guide people unconsciously. Values are internal balance to discern what is good and bad. They are hierarchical and some of the are more important than other depending what social group you are looking at. **Basic Human Values**, derived from the studies of Shalom Schwartz [2], are the following shown in 5.1:

- Self-direction (SD): independent and outside of control of others, needs of control and mastery your own life.
- Stimulation (ST): in search of excitement and thrills, novelty and challenge
- Hedonism: (HE): enjoy and sensuous gratification for oneself
- Achievement (AC): setting goals and then achieve them, personal success and necessity to demonstrate it
- Power (PO): dominate other and control over resources, necessity of a dominance/submission class status where you can dominate others.
- Security (SE): security, health and safety in society
- Conformity (CO): obey clear rules and structure, self restraint, avoid initiative that can disrupt the group

- Tradition (TR): practice of the past so customary
- Benevolence (BE): help other and pursuit of general welfare, organismic need for affiliation
- Universalism (UN): social justice and tolerance for all There were also an initial referral to Spirituality, but later abandoned just because not enough universal among extremely differentiated cultures.

They are grouped for simplicity and internal coherence:

- Self-Enhancement (SC1): (Achievement, Power, Hedonism)
- Openness to Change (SC2): (Stimulation, Self direction, Hedonism)
- Self-Transcendence (SC3): (Universalism, Benevolence)
- Conservation (SC4): (Security, Tradition, Conformity)

We decide to include Schwartz's discoveries in our job due to the engine, behind them, that pushes communities to birth and share values. This is fundamental in order to discover which levers best perform in pushing other human beings to action inside groups (also social media ones). Another important reason is they are universal, as said, they originate from most basic biological needs: socialization, survival and welfare.

5.3 Five Factor Model

The Five Factor Model defines how to retrieve and calculate, on a one to five scale, each of the five traits by a likert questionnaire [20] so asking people how similar their behaviour are with respect to the written situation. The Five Factors are:

- *Openness* is a measure of how a person embrace new experience in a trial of avoiding habits and routine. How a person is able to exit his comfort zone and engage risky behaviour. Usually these people are creative and tend forward art and humanistic path.
- *Conscientiousness* is an indicator of how much a person is organized and efficient. If she is self-disciplined and can reach goals fixed in time with a well defined path with high concentration and focus.

- *Extraversion* means that a person is explicit and not shy in external its thought and behaviour, a person who is solar and empathetic with others. He is attention seeking and in search of company of others if high in this value. People in this high category tend to be dominant in a group.
- *Agreeableness* people are compassionate and friendly, cooperative and gregarious. They are often seen as naive or submissive.
- *Neuroticism* is the most difficult values to detect because of its own dramatic nature. Person high in neuroticism experiences unpleasant emotions easily, such as anger, anxiety, depression, and vulnerability.

Agreeableness, Conscientiousness and Neuroticism are considered part of the socialization process, in fact if we correlate the output of being a micro influencer we find that the first two are strictly correlated: tendency of gaining group consensus and determination in a constant behaviour in posting tweets in our case, for the last one instead it is extremely difficult to detect neuroticism for the tendency of people in hiding personal problem when sharing with the public. Extraversion and Openness are more related with self-actualization process. All these traits are precious for detecting anomalies in behaviour or just highlight great capabilities of individuals. The FFM was used by Cambridge Analytica during the elections in USA and this drove to an international scandal on abuse of human privacy and data utilization. Five Factor are not just yes/no features, instead in base of the result of the analysis they gave a scale of values with which is possible to derive many information about personality. There are many thoughts about the legitimate use of FFM, many critics have been raised, in particular in the possibilities of these outputs manipulation in order to create minorities among communities and discriminate groups with certain score. Many researchers are raising increasing doubts on the ethic of these data usage, as an example Kate Crawford said that there is no reason to stop analyzing data, instead we have to acknowledge which is the goal we are pursuing with this data, but most important we have to understand which direction we want to push the society to and create later the technology to reach it and not vice versa [22]. With this in mind the exploitation of big5 is straight forward and must be based on transparency and clear definition to final user of which are the real reasons why we are analyzing this possible sensible data. Working with individuals and groups we tried to exploit this two set of parameters in a measure that Big5 is related to a person as an individual meanwhile Schwartz Personality Values as a person projected in a community field so common/shared social scale.

Chapter 6

How text can be a valuable source for measuring personality traits and social indicators

6.1 Personality from survey

Psychologists usually determine personality through questionnaire, which give life experience examples to agree with or not. Questions are directly connected with traits. Each answer contributes to increase or decrease score. At the end of the survey there is an aggregations to determine each trait in a range 1-5. NEO-PI-R from McRae and Costa [20] is a questionnaire with 240 questions, need about 45 minutes to complete it and is used to retrieve Big5[18]. It has been said that self report questionnaires are subject to self representation and subjectivity in answering, because is the patient himself who look at real life filtering it through is eyes. The philosopher Schopenhauer talked about Maya's Veil so that a person through her experiences of life always seen what occurs as self centered. We can bring the example of a person born wearing light blue lenses glasses not aware of bringing it and he spends his whole life wearing it, he doesn't know that many colors exists in reality and will tell others that real world is like he sees it. Unluckily the self report problem is embedded in answering questionnaire. Some attenuations of the problem could be brought by asking similar questions to friends, family members and foreigners on what they thought about the subject in analysis, this situation mitigates the problem but it doesn't solve it for the same subjectivity reason. As the number of individuals delineate the charac-

| | | | | | | |
|-----|-----------------------------------|----------------------|----------------------------|--------------------------------------|--------------------------|--------------------|
| 1. | Worry about things. | Very Inaccurate ● | Moderately Inaccurate ● | Neither Accurate Nor Inaccurate ● | Moderately Accurate ● | Very Accurate ● |
| 2. | Make friends easily. | Very Inaccurate ● | Moderately Inaccurate ● | Neither Accurate Nor Inaccurate ● | Moderately Accurate ● | Very Accurate ● |
| 3. | Have a vivid imagination. | Very Inaccurate ● | Moderately Inaccurate ● | Neither Accurate Nor Inaccurate ● | Moderately Accurate ● | Very Accurate ● |
| 4. | Trust others. | Very Inaccurate ● | Moderately Inaccurate ● | Neither Accurate Nor Inaccurate ● | Moderately Accurate ● | Very Accurate ● |
| 5. | Complete tasks successfully. | Very Inaccurate ● | Moderately Inaccurate ● | Neither Accurate Nor Inaccurate ● | Moderately Accurate ● | Very Accurate ● |
| 6. | Get angry easily. | Very Inaccurate ● | Moderately Inaccurate ● | Neither Accurate Nor Inaccurate ● | Moderately Accurate ● | Very Accurate ● |
| 7. | Love large parties. | Very Inaccurate ● | Moderately Inaccurate ● | Neither Accurate Nor Inaccurate ● | Moderately Accurate ● | Very Accurate ● |
| 8. | Believe in the importance of art. | Very Inaccurate ● | Moderately Inaccurate ● | Neither Accurate Nor Inaccurate ● | Moderately Accurate ● | Very Accurate ● |
| 9. | Would never cheat on my taxes. | Very Inaccurate ● | Moderately Inaccurate ● | Neither Accurate Nor Inaccurate ● | Moderately Accurate ● | Very Accurate ● |
| 10. | Like order. | Very Inaccurate ● | Moderately Inaccurate ● | Neither Accurate Nor Inaccurate ● | Moderately Accurate ● | Very Accurate ● |

Figure 6.1: Example of questions from NEO-IPIP questionnaire

teristics of the subject increases, the mean enforce or destroy the definition of the subject characteristics. As describe in an article [46] if we collaborate with pool of experts in psychology to better define a common definition of personality over a person, we can force at least the same input same output through automated analysis, to do so the lexical hypothesis runs in our help. It is clear to see that we are driving psychology on the road of engineering, so in contrast with what said in incipit by Galimberti, but we are also considering Brian Little in his Ted Talk when we divide the subjectivity of a person from his categorization due to his traits. We know that we are not a bunch of numbers , as individuals what define us is our personal experience, knowledge, sentiments, and every individual is extremely different from other even if genetically similar, but here we are considering the more categorical and general aspect of human personality, the ones that allow us to describe a user belongs to a particular groups definition, preserving him as a standalone case.

On the other side we explore Basic Human Values. The Schwartz Values Survey counts 57 items and takes about 20 minutes to answer with the previous approach but to compute Basic Human Values[39],[2]. We described this values former in the text, but we want to highlight why we choose also this model to define our users. This study conducted by Shalom Schwartz reason more over cultures and communities than the Five Factor one. In this sense he defines cross cultural values in a way to even categorize people in a

higher level so defining national scores (it has been done¹). We exploit this reasoning to shift the attention of the user performing the questionnaire to the community in which he lives, and doing so we can better avoid subjectivity thanks to the higher number of user involved in a particular reality, obviously they can influence each other following the leaders opinion, but also this phenomenon is of interest in our research. Schwartz has understood that putting in a priority order what are more likeable values, defining values as described in the previous chapter, is an alternative way of clustering users in communities and virtually accumulating people with same principles in mind. The Schwartz's work can build the bases of social problem solutions: as an example of this we can settle immigrants in community better predefined in acceptance of other and less worried about losing traditions, or instead put victims of abuses in a community where he feels protected by the common predilection of security, but the scenarios are a lot; let us think about borders of nation or in trade routes where cross cultural economics found common base in creating relationships leveraging on more desirable traits of both community. We decide to introduce Basic Human Values to detect community pattern in following leader/influencer reasoning about reality, in this way much more work can be performed in analyzing also the followers traits pattern.

The last approach derives from Carl Jung studies and it is also parallel to Five Factor Model, and we do study it but we do not include it. The 16 personalities factor interview is derived from Jung studies [56] can be performed online² and is connected to personalities typologies in the study of Myers and Briggs [38]. The test seems very accurate and it is, in fact it is a clear example of what discussed in the previous section, here we can describe a person as belonging to a certain category but without going deep in personal experience. As illustrated in figure 6.2 the characteristics of most of micro influencer is ENFJ - the protagonist described as leaders and charismatics as in fact they are. We retrieve this output also from writings passing by Big5 Traits. We decide to not include 16pf in our work because of its redundancy as compared with Big5 from where we derive them.

6.2 Personality from social media

Social media platforms such as Twitter and Facebook are source of people writings which can be used to retrieve personality traits following *lexical hypothesis*[16]. Many studies [36], [37] demonstrate the model validity and

¹<https://www.europeansocialsurvey.org/data/themes.html?t=values>

²<https://www.16personalities.com>

ENFJ PERSONALITY (“THE PROTAGONIST”)

Everything you do right now ripples outward and affects everyone. Your posture can shine your heart or transmit anxiety. Your breath can radiate love or muddy the room in depression. Your glance can awaken joy. Your words can inspire freedom. Your every act can open hearts and minds.

— David Deida

ENFJs are natural-born leaders, full of passion and charisma. Forming around two percent of the population, they are oftentimes our politicians, our coaches and our teachers, reaching out and inspiring others to achieve and to do good in the world. With a natural confidence that begets influence, ENFJs take a great deal of pride and joy in guiding others to work together to improve themselves and their community.



Figure 6.2: ENFJ personality output from [58]



Figure 6.3: Principal Social Media that could be useful for our tool

also illustrate how spontaneous or not social media posts are. We are aware we deal with human being and so it is important to keep in mind what Brian Little [53] says about typologies. Human typologies are the objective characteristics of an individual, but the most important for human being are the implicit or free aspects of personality, the ones that make ourselves unique, so the goal we have in lives. On the other side Five Factor Model, Basic Human Values and word embeddings[21] (later detailed) are useful for grouping similar tendencies in human behaviour and extract information. Analysis of parameters and values are conducted on candidates' writings over long periods of time, so evolving a chronicle of the subject, but at the same time parameters are extremely influenced by contemporaneity and by the actual moment. These two aspects allows to calibrate the weight of time. Deriving information from social media can mitigate the filter a person impose on himself when self reporting. We can download information and users writings in a massive way. Someone can say that also in social media we expose just a frame of our personality, usually the kindest one, but this isn't true on large scale in fact studies[40] declare how average user tends to be spontaneous in writing post without over thinking too much on what she is posting. There are at the same time other restriction on data retrieval from social media platform and in particular from Twitter we use in our work: lenght limitation on what to post in a tweet, handle, hashtag depauper the possibility of better explanation of a person thought. In this way is promising the work done by authors of Empath³ [42] in which the base group of words retrieval is a site where users publish book. In this sense the lexicon is much wider and the feeling are much better explained. We reason about applying our study in other application: one of the could absolutely be to find the book or the authors that match your values and your personality traits. This could be a powerful tool both for authors and editors to find their public and early understand if certain contents would be appreciate.

6.3 How to avoid Panopticon?

Scientific researches and consequently applications can leads to amazing improvement in life style granting new benefits, overcoming problems, reduce disequality, or can lead to the opposite situation if applied as a non ethical tool to gain unfair advantage to other damaging them. In our study we want to suggest a fair use approach to avoid the Panopticon situation of the worst distopia. A Panopticon is a society where the government has total

³<https://github.com/Ejhfast/empath-client>

surveillance of citizens, it is described in the romance "1984" by George Orwell. But why our tool is prone to that dramatic evolution? Imagine that all social platforms users are analyzed to find micro influencer hints, but both writings and sensible data are collected in the process, we will give unlimited access of potential client data to companies and in such environment users will be subjected to minority assimilation and groups denigration. We, as scientific people, cannot justify ourselves with the phrase "I am just an engineer" and shift the focus on the mathematical error instead of the approach used. Every scientific discoveries, and in particular this one, has a huge impact on the life of people in real life. Some people can be designed as not suited for a job or an assurance due to his psychological parameters or for what he said in an ironical situation. At the moment machines can't detect when a person is sarcastic or use certain words for the purpose of information and not for terrorism. We must be aware of not give all the decisional power without supervision to automatic processes that are not capable of detecting this exceptions. On the contrary there will be situation in which children are detected as members of a criminal gang, or life assurance is not elapsed on people with particular strange scores in neuroticism. It has been scandalous both homosexual detection and racial detection from face recognition cases. Jacquard developed a list of questions to answer before publishing and also before developing a particular scientific discoveries. They are questions about social and environmental impacts, if our discovery will delete jobs places and increase disoccupation.

Chapter 7

Word embeddings

We exploit word embedding when we want to extract significant information in text analysis. Computers are not aware of what a word means, at least in the way humans are. Machines are able to understand signal and consequentially numbers. This dilemma is at the center of actual studies on Natural Language Processing and it's at the moment one of the highest obstacles to overcome. This means that machines cannot understand sarcasm, nor the explicit meaning of a phrase, nor deduce the intention of an assertion. Word embedding comes to help towards machine understanding of syntactic and semantic values of words. As said, word embedding transforms words into real numbers arrays. Embedding terms is referred to mathematical transformation of an object into another through a specific function. There are many approaches to this transformation but word embedding is capable of reducing the dimensionality of the problem preserving the original meaning of words and their related environment of research. A technique is called TF-IDF so term frequency-inverse document frequency in which each word is counted whenever it occurs but its presence is weighted due to the fact that it is a frequently used word in every document instead of a rare one, so giving more strength to a rare one that can possibly be a characteristic of that document. Here the computational used is one-hot encoded so we have a huge vector representing the whole vocabulary and each array is just set to one the number at the indexed word position. Here words with similar meaning have similar indexes set to one. First instances of this approach date back to the 60s and we have track of them in the so called Vector Space Model. A vector space of dimension N is described by a set of N linearly independent basis vectors. For text processing, each one-hot encoded word is a basis vector of the space, so that its dimension corresponds to the total number of distinct words in the corpus. A group of words, which we refer to as a document, can be represented as a vector V in the N -dimensional space as a combination

of the basis vectors. Depending on the technique used for combining the words, vectors will have different values, hence different representations in the space. Possible methods for creating document vectors are binary weights and TF-IDF.

7.1 Comparison on techniques

Semantic vector space model is a method of representing words with vectors containing real value numbers. These values can be exploited differently depending which application you are using them for. Many methods use as similarity parameters the distance or the angle between two words/vectors spatially oriented. A more detailed analysis will consider not only the whole vector as a unique element, instead it will compare distances among particular values of these vectors, keeping in mind what is the objective of the clustering and similarities searched. The last is the concept of multi-clustering. Subdivision in learning word vector:

Global matrix factorization i.e. Latent Semantic Analysis: good statistically, bad clustering and analogy.

- LSA term-document
- HAL term-term
- COALS solves HAL numerical occurrences problem
- PPMI
- HPCA

Local context window i.e. skip-gram model: good clustering and analogy, bad statistically

- Simple neural network
- Skip-gram
- Continuous bag of words (CBOW) inner products between vectors
- iv LBL

Analysis of model properties to achieve linear meaning directions and give an explanation of how this can be reached through global log bilinear regression models. Glove adopts a weighted least squares model over word-word co-occurrence counts to handle both previous benefits of gmf and lcw. Primary source of informations in unsupervised learning is statistical data on words occurrences in a text corpus.

Chapter 8

How information propagates itself in a community

Another pillar of this paper is the understanding of human interaction and if there are powerful nodes in a social graph and among communities. [52] In the context of social graph we can cluster members having similar friendships in regions. It emerges a structure where some node acts like intersection between two subsystems highlighting ties. There are two types of them:

- *strong ties* represent close and frequent social contacts and tend to be embedded in tightly-linked regions of the network.
- *weak ties* represent more casual and distinct social contacts and tend to cross between these regions.

The way these people interact with shows us that contamination of behaviours, beliefs and thoughts is faster inside a region full of strong ties, but slower to contaminate other dense regions of the social graph. Reading the book of Easeley and Kleinberg [52] the authors take in consideration the possibility of the human tendency to conform, but underneath the surface what push choice and action is the information each member possesses and that has previously used to take a certain decision. The trigger to action inside a social community derives both from personal information and from other users we consider having important knowledge on the field. Bring this reasoning to the extreme case shows us the abandon of personal information and follow the crowd and the allowing the information cascades as shown in 8.1. We use this discoveries later in this paper to compute the embeddedness score which computes the overlapping of followers among different communities to say if the influencer action can be effective or not. The idea is that if two or more users with influencer characteristics talk about same topic with

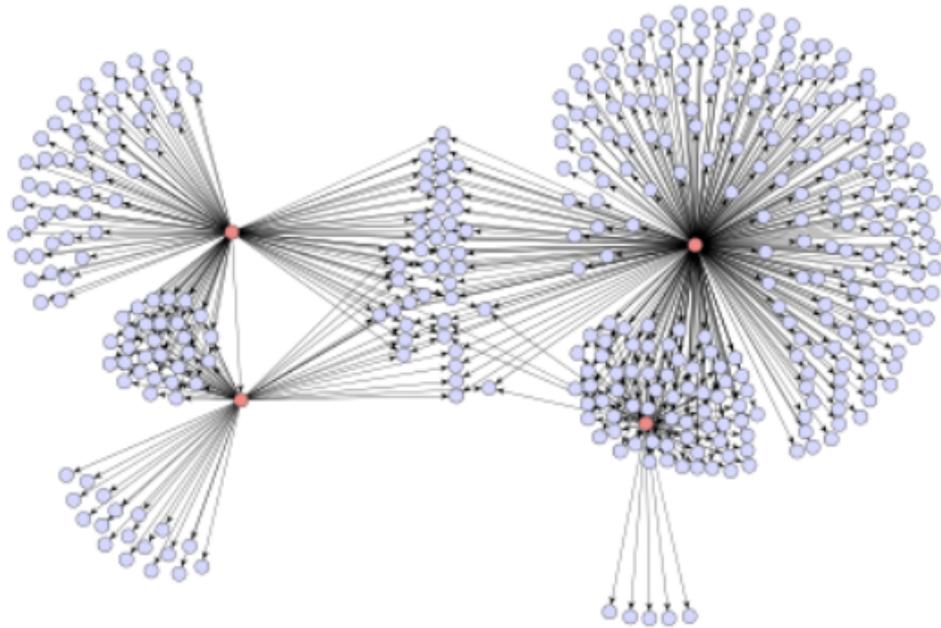


Figure 8.1: Information Spreading in social graph from [35]

similar opinions so the information inside the same community will spread rapidly.

8.1 Groups Behaviour

In the last CIKM2018 we follow a seminar explaining the group behaviour in social network, this gave us the possibility to explore how our work consider these dynamics [45]. We reason about the needs of understand community interaction with respect to its micro influencer: certainly they share interests, values (Schwartz), but they do not have same Big5 personality traits, because the micro influencer is the more active one in the relationships and so overcoming others impact on the community. Is there a one direction communication? Or is it different from standard influencers top bottom situation? Are micro influencer followers more prone to acquire high quality products and with more awareness in action? Are followers less far from their micro influencer in terms of expertise? All these question can be answered with continuing working in the path cited [45].

8.2 Networks, Crowds, Markets

All this section will refer to aspects treated in the books [52]. In this egregious work of Easley and Kleinberg we find the soul of our work, it is the place of many of our intuition find their validation and comparison. Why is it so coherent with our work? Because starting from the title and running through all its 819 pages we can see experiments and results of topology and social graph investigations and so relationships with the market environment, technology spreading, social influence so everything matching our work. One of the concept we exploit from the books is the **homophily**. It indicates the tendency of sub groups creation inside a greater community among people sharing common interest, but also ethnical, racial and geographical traits. It emerges in drawing of social graphs that users tends to stay closer to individuals that are similar to them. If this concept enforce the diffusion of certain behaviours inside the sub group, on the contrary the possibility of cross groups contamination is compromised. Let us consider this term with respect to the micro influencer environment: high homophily will increase engagement in the relationship micor influencer followers and that is what we are searching, we take in account this passage through the recall score that monitor how much engagement there is between them. High homophily is good for micro influencer but less for macro influencer who constantly try to expand the number of followers. High numbers are achieved by chosing many micro influencers following this thoughts, so homophily enhance our engagement power and there is no problem in spreading of information because it is not needed with this approach outside of the community. Now we can focus our attention to the concept of **passive engagement**: the tendency of a user on maintain a passive behaviour with respect to people sharing their opinions. In a relationships, in particular in a social graph, there are mutual relationships so active conversation and engagement from both direction or mono directional one where an end is active and the other is passive, it has been notice that when the community enlarges itself the numbers of passive relationships increase because of the impossibility of maintaing so many relationships in an active way as shown in figure 8.2. Again we can now consider the passive engagement in or context: it is clear that we are going in the right direction with micro influencing because if larger community lose reciprocal interaction so the engagement between the two part too. We look at micro influencers who can push to action their followers but obtaining their trust and not in a passive way, but sharing infromation and knowledge as in real life context. We store this result in the engagement score.

Let us consider the **cascading behaviour** also called *social contagion*: this concept express both spreading of epidemics, technologies and social

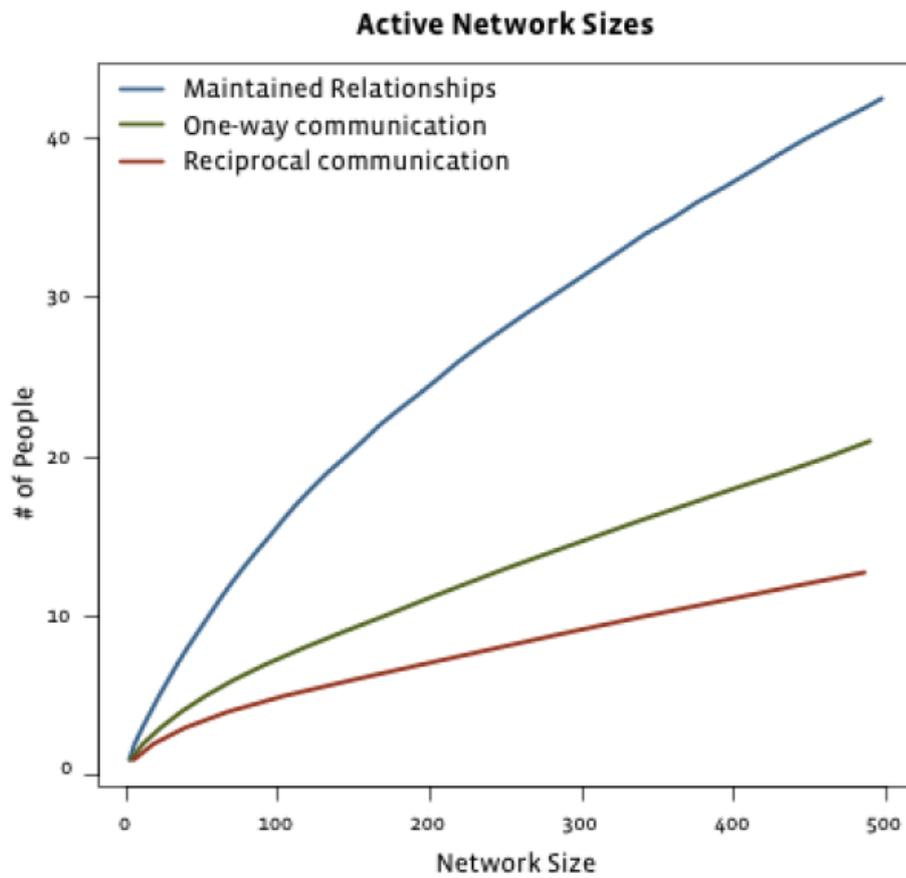


Figure 8.2: The number of links corresponding to maintained relationships, one-way communication, and reciprocal communication as a function of the total neighborhood size for users on Facebook. Image from [60]

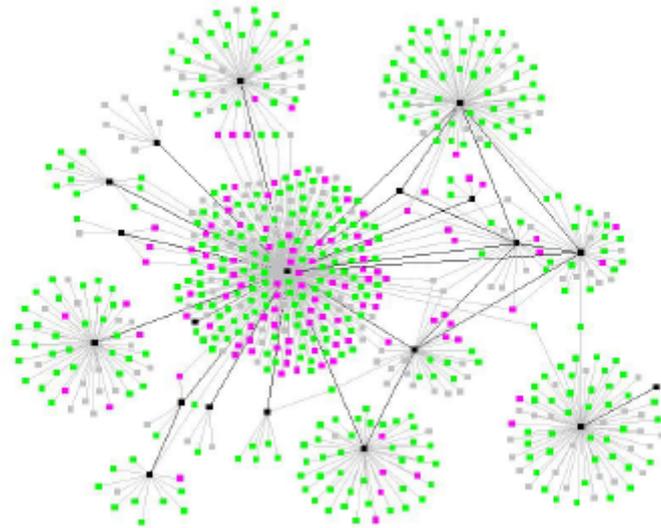


Figure 8.3: The spread of an epidemic disease (such as the tuberculosis outbreak shown here) is another form of cascading behavior in a network. The similarities and contrasts between biological and social contagion lead to interesting research questions. (Image from Andre et al. [61].)

influence.

When individuals have incentives to adopt the behavior of their neighbors in the network, we can get cascading effects, where a new behavior starts with a small set of initial adopters, and then spreads radially outward through the network. [...] We will also find that the diffusion of technologies can be blocked by the boundary of a densely-connected cluster in the network a closed community of individuals who have a high amount of linkage among themselves, and hence are resistant to outside influences. [...] social contagion tends to involve decision-making on the part of the affected individuals, whereas biological contagion is based on the chance of catching a disease-causing pathogen through contact with another individual.

Also in this case we consider it in the micro influencer scenario and mathematically compute it with embeddedness score saying that if common followers of multiple micro influencers are affected by the same "pathology" the morbus is fast spreading.



Figure 8.4: Cascading adoption of a new technology or service (in this case, the socialnetworking site MySpace in 2005-2006) can be the result of individual incentives to use the most widespread technology either based on the informational effects of seeing many other people adopt the technology, or the direct benefits of adopting what many others are already using. (Image from Google Trends, <http://www.google.com/trends?q=myspace>)

Chapter 9

Other relevant studies on this field and obtained results

9.1 Studies on influencers figure

Nowadays many studies related on measuring influencing abilities on social media have been performed, but most of them are focused just on quantitative data stored on platform like counters [1]. It is the case of number of followers, retweet, comments, etc. We used these studies as base ground for the first part of our work. The other relevant aspect of similar studies is that they look at influencer in general meanwhile we shrink the attention to the emerging position of micro influencer [23]. On these studies emerge that using statics to investigate which social media user is more effective than other in spreading informations result convenient and effective. Companies are interested in these results mainly to save money in advertising company reaching directly influencers with high engagement with their public and community. In the work of Anger and Kittl [1] as example the tried to develop a score system calle SNP (Social Network Potential) in the pursuit of correlation between social statistics between most followed influencers in Austria. On the other side in the work of Bakshy et al. [23] they determined how influence analysis can directly impact marketing investments. We have analysed also the important aspect of merging personality data retrieval from social media post and in this area one of the papers of reference is the one performed by Bigonha et al. [24], where they deeply go in search not just of user talking about a certain topic but understand if they are detractor or evangelists.

9.2 YouTube experiment at IDIAP

On the wave of enthusiasm and spirit of collaboration we contacted the Idiap research center at EPL in Losanne and we obtained their dataset to investigate their brilliant job on vlog over YouTube analysis[31], [32], [33]. As said more and more information are vehiculated through non verbal or textual methods, so it is precious their work for future developments of our tool. Their approach inspects audio video feature and transcription of words said in vlogs. At the same time they used automated analysis such as image details recognition in high spike of gray shade and face expression understanding, meanwhile for what concern audio features the associate spike in voice with anxiety or excitement, finally correlating the whole process with meaning and use of word for semantic and syntactic analysis of personality. But another reason why it is useful as a model to reproduce is that they create their gold standard with Amazon Mechanical Turk, a payment service offered by Amazon where people are paid for human task validation. In particular they ask to so called Turkers to watch videos and answer a questionnaire in order to label the vlog itself. In this questionnaire ask about personality of vloggers and other useful resources. We, instead cannot validate our model through AMT due to costs and disposal time in the thesis period.

Part III

Early Detection of Micro Influencers

Chapter 10

Mathematical derivations

In order to assess if a user is a micro influencer we need a scoring system. The first step is to select Twitter users who have recently written about the searched topic, i.e. using the topic as hashtag in their tweets. Then we filter out user having less than 1k and more than 100k followers according to Forbes definition¹ of micro influencer. The next step is to retrieve whole followers of each user in list. Finally we download all tweets in users history. At this point we can compute the first out of three score.

Embeddedness score is derived from Easley et al. book[52] where the authors talk about *neighborhood overlap*. The concept expressed is information spread fastly inside a community where at least two influencing members speak about same concept. In our work we derived that if two potential micro influencers have almost the same followers group, they belong to the same community. In mathematical words we count how many times each follower of a potential micro influencer appears in the other candidates' followers list. For each user we sum up followers repetitions and then divide by number of followers of the candidates. If the mean number of time they appear is greater or equal 1, this means that at least each one of his followers appears in another followers list. This realizes the concept of *neighborhood overlap*.

$$\frac{\sum_{i=1, j=1, i \neq j}^{n, k} same_follower_{i, j}}{m} \quad (10.1)$$

In 10.1 n represents number of user i followers, while k represents number of potential influencers in that specific topic, m stands for number of user followers we are computing score for.

¹<https://www.forbes.com/sites/barrettwissman/2018/03/02/micro-influencers-the-marketing-force-of-the-future>

Recall score is computed selecting for each significant tweet (so the ones containing the hashtag selected) how many retweets are performed by follower. We take inspiration from *Interactor Ratio* of Anger and Kittl [1]. It is important to notice that we do not just count how many retweets each tweets received, instead we check how many are performed by follower. This concept is derived from the definition of micro influencer ² in which is highlighted the high engagement with his community so his followers so we are not interested in people that can retweet but do not follow. In particular Twitter API allow the collection of just 100 retweet, because the number can change instantly and you need stream but we do not want to wait indefinitely. So we worked percentually over the last 100 retweets over each tweet posted by the potential micro influencer.

$$\frac{\sum_{i=1}^n \frac{RetweetByFollower_i}{TotalFollowers_i}}{SignificantTweets} \quad (10.2)$$

Interest is the third score and it is derived from our reasoning about how much a user posts tweets about the specific topic in analysis. It is computed counting inside how many tweets the word used as hashtag appears, divided by the total amount of tweets analyzed for that specific user. This score is useful to understand if a user can be considered passionate of a topic or if he tweets about it rarely.

$$\sum_{i=1}^n \frac{SignificantTweets_i}{TotalTweets_i} \quad (10.3)$$

Engagement is an extra score that we elaborate from the tool called Grin³. We modify it with no enterprise parameters callable by twitter rest api. We measure the engagement of public with the micro influencer summing up number of like, number of retweet and than divide this sum by number of follower this is further divided by total number of tweets so analyzed.

$$\frac{\sum_{i=1}^n \frac{Likes_i + Retweets_i}{Followers_i}}{SignificantTweets} \quad (10.4)$$

Schwartz is a multiple score composed by ten values. We load an already trained word embeddings model from GloVe⁴ then we considered words given by Schwartz in his studies over Basic Human Values [2] and we retrieve their

²<https://www.grin.co/blog/the-ultimate-guide-to-micro-influencers>

³www.grin.co

⁴<https://nlp.stanford.edu/projects/glove/>

embeddings. For each human value, out of ten, we compute the centroid from the example words and we use them as multidimensional points in space to compute euclidean distances of all other words used in tweets by a user to put them in one of the ten values. We compute the centroids of all words in each one of the ten values and store the number of words for each human values. We lastly compute the score by multiplying number of words for $1/\text{distance}$, where distance is the mean centroids of all words used by user in that semantic area from the original reference centroids given by sample words by Schwartz[41]. Before adopting the just explained strategy on Basic Human Values, we explore the empath tool⁵[42]. It is useful to better understand the mechanics behind embeddings and the topic understanding on large scale text.

Big5 score is composed of 5 different score: *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, *Neuroticism*[3]. We described each of them in the first part of this paper. We follow the work of Giulio Carducci to compute them in the case of our dataset[43]. We follow his choice in the selection of FastText⁶ dataset based on Wikipedia embeddings. We exploit the strategy of user wise computation of tweets so analyzing the whole history of user as a single big text, both for time consuming reason and for a more heterogeneity in score. The other path available is computing Big5 values per tweet and later make a mean for each values over whole tweet corpus of the user. We choose the first way after performing both because the large corpus, about 3000 tweets per user has the tendency to uniform results on the mean averaging scores, instead we want to preserve as much as possible the diversity among users. We consider each tweet has its own topic and user mood so we want to preserve the overall changing of the user in his whole story. It is interesting to notice that all user analyzed having followers count in the range previously cited has the Jungian type ENFJ that is "the protagonist"⁷. You can find the process of train model and tokenization and file cleaning at the github repository⁸

⁵<https://github.com/Ejhfast/empath-client>

⁶<https://fasttext.cc/docs/en/english-vectors.html>

⁷<https://www.16personalities.com/enfj-personality>

⁸<https://github.com/D2KLab/twitpersonality/tree/master/test>

Chapter 11

Implementation details

11.1 Technical choices

We used Twitter platform to create our datasets downloading tweets of user. We chose that one thanks to already existing interfaces with the platform and the extended community of developers working on it. We use the *Tweepy* library which works with python and Twitter platform in a light and fast way, allowing REST API request translation as normal Python functions. We can now consider all the steps configured in the infographic [12.1](#):

- **Authentication:** Twitter platform has really strict procedures in authorization process, so we have to create our developer account on the site in which we explain what are our purposes with this application and what kind of data we want to retrieve, we have also to clarify if we are a private, public, or government user and why we need these data. Later on we have to retrieve token both for application and to have an unique identifier of the user who is using it [11.1](#). Once this passage is done we can store both key and token given by twitter platform as string in a local file to avoid re performing authentication token retrieval each time we use the tool.
- **Users Retrieval:** once authenticated we ask Twitter app to perform a search query with a topic/hashtag given in input. Twitter answers with a list of tweets randomly chosen but having that hashtag included. From that lists of tweets we find users posting those tweets and because of the possibility of duplicate user we filter them out.
- **Followers:** We now have users candidate to be micro influencers, the next step is to download for each user a huge array of identifiers that



Sign up for Twitter >

Authorize micro_influencer_psy to use your account?



micro_influencer_psy

www.microinflpsy.com

This app will select just Tweets with a specific topic, from selected user and their retweet, trying to understand if the user can be a micro influencer of that topic based on a metric.

Remember me · [Forgot password?](#)

Authorize app

Cancel

This application will be able to:

- Read Tweets from your timeline.
- See who you follow.

Will not be able to:

- Follow new people.
- Update your profile.

Figure 11.1: Authorize Twitter Application Micro Influencer detector to access through your account

define how many and who are the followers of each user in analysis. We store them in a csv file for later usage.

- **Tweets:** If users and followers are the base with which we compute marketing score, tweets posted by users are fundamental for writings analysis and Big5 and Basic Human Values computation. We, as before, start from the list of users and for each of them we download whole tweets posted from her initial subscription to Twitter and we clean them eliminating strange spaces, after that we store them in a tab separated value containing id of each tweet and its connected text. The id is important for counting how many retweets and like are related to it.
- **Recall:** The concept of recall score is later explained, here we just say that recall is an extreme time consuming task because we are asking for each tweet from who it is performed and then we check if it is a follower of the user or not. These continuous and persistent interrogations at Twitter API trigger a Rate Limit Error exception that would stop our application if we do not handle it properly. We analyse the exception and find that it forbids new requests from the same application and user for the following 15 minutes, so we create a sleep function in python that consents to the application to stay in an idle state until it can request again data. This step requires days.
- **Engagement:** This step on the opposite situation is computed straight forward after followers retrieval because we check how many followers of a user repeat themselves in other potential micro influencers' followers list.
- **Interest:** Here we count in how many posts a user used the specific hashtag we search. In here we want to exclude users who just talked about that topic a few times.
- **Define Output:** Output is defined on the premise of Embedness score as later defined and its threshold and then if user score of interest and recall are higher than the average among all other potential micro influencers in the list we retrieved.
- **Big5:** this step recreates the environment of Giulio Carducci thesis and its tool¹. So we use FastText dataset and then we used their pre-trained word embeddings and later we move on SVM already tuned by Carducci to exploit its potential in analyzing tweets. We use the user

¹<https://github.com/D2KLab/twitpersonality/tree/master/test>

wide tool instead of the tweet wide as explained on the Five Factor Model Section.

- **Basic Human Values:** this step and the Big5 one can be performed in parallel, after tweets download tweets, with marketing score computation. To save time in order to speed up the process, if you have enough cpu resources you can perform these three main chunks of work together. This computation of Schwartz values is made from scratch without barely computing values as suggested in the overview [2]. We load the GloVe pre-trained word embeddings and we cluster examples word to find centroids, later we scan every word in tweets and assigned it in a specific Basic Human pool. Then we compute score with distances and number of words.
- **Create Table:** in this step we collect all previous results and store in an matrix like table starting from marketing scores, trough personality traits and concluding with the output.
- **Clean Table:** in this phase we clean missing or outliers values refreshing the table just with exploitable values. could happen that we not find values in a certain Schwartz values or that the centroid is exactly in the same position of the reference one, or that some error misplace some dirty values.
- **SVM, Random Forest, CNN:** here we created three different tools to perform prediction, as conceivable by looking at figure 11.3, the SVM is the best performing one. The reason behind this choice is in the numbers of data retrieved, in particular in number of users retrieved and in large part to the fact that just a decimal part of all user is considered as micro influencer as you can see in the pie chart 12.5. If you look at the graphic it appears that traditional machine learning approach such Support Vector Machine perform better than the Neural Networks one with low number of samples but later they will not improve anymore, fastly saturating. At the moment we are writing the thesis the experiment is still going on, in fact we lately understood that the definition of threshold were not the problem, instead the scarcity of positive examples in the dataset determines the low score in table. 12.2.

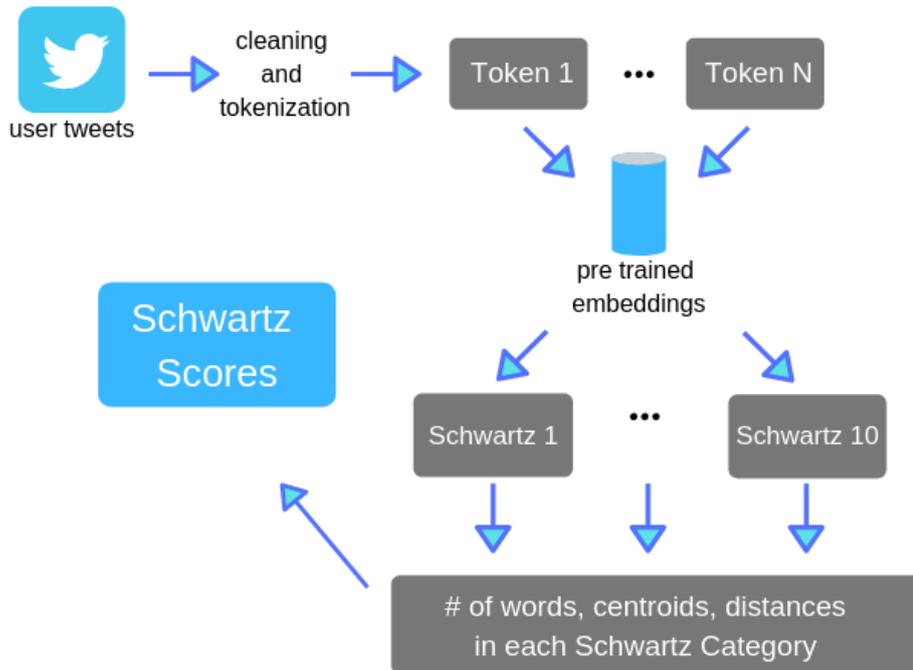


Figure 11.2: From user tweets to Schwartz scores

11.2 Packages and libraries

In the implementation of the code we use python3.7.6 in whole code written with the support of Pandas, Numpy, Scikit learn, Gensim, Nltk and many other machine learning related libraries. They are fundamental in every phase of a data science analysis. For what concern instead CNN we both experiment Keras and pyTorch to calibrate the convolutional network. We use cross-val-report, cross-val-score and scoring system given by scikit learn to determine our output and at the same time, thanks to the binary case scenario we used also visualization of the confusion matrix to understand where are the problems in the tool.

11.3 Tweets Preprocessing

We use the natural language toolkit² to perform data cleaning. In a first step, immediately after downloading user tweets, we store them as a tab separated

²<https://www.nltk.org/>

value file removing all newline and tabulation in original text. In a second phase we perform the following procedure:

- **stop-word removal** "Uhm, where is the leader? @johnsmith #officelife. :)" to ", where leader? @johnsmith #officelife. :)"
- **punctuation removal** ",where leader? @johnsmith #officelife :)" to "where leader @johnsmith #officelife :)"
- **emoticon removal** "where leader @johnsmith #officelife :)" to "where leader @johnsmith #officelife"
- **handle and url removal** "where leader @johnsmith #officelife http://.../" to "where leader #officelife"

Once cleaned, the text is tokenized on a space based level and each token analyzed searched in the corresponding trained embeddings vocabulary. The word embeddings vocabulary is made of 300 dimensions vectors that has been demonstrated by Landauer and Dumais as the optimal number of features for distributed word representations[44]. When ready, each word in user tweets is searched in the GloVe dictionary and assigned to the closer (Euclidean Distance) Schwartz centroid.

11.4 Learning model choice

We develop two models: the first is a topic related one, with which a company looking for micro influencers gives us a specific research area so we train the model with values of users posting that topic; the other one is created with the aim of reaching a general definition of micro influencer without specialization. We perform stratified kfold with Support Vector Clustering(SVC), Random Forest Classifier and Convolutional Neural Network. We finally choose the first one after looking at results in terms of validation metrics as shown in table 12.2. We use both Big5 and Basic Human values to train the model, due to different scales we formerly scale to a range 0 to 1 the values.

11.5 Why we choose SVM

A **SVC** is a particular case of Support Vector Machine that uses a rbf (radial basis function) for kernel described in figure 11.1, where x and x' represent two features of the model. SVM are used in supervised learning when we want to create an hyperplane that maximize margin between two calsses (in

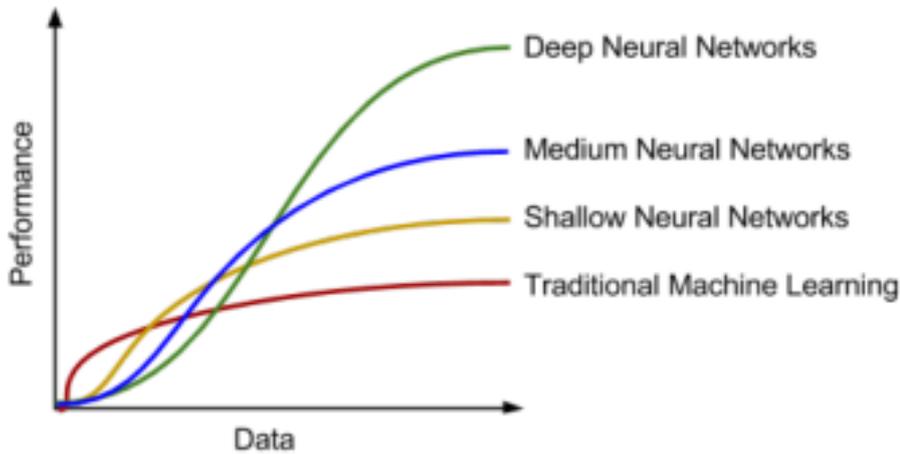


Figure 11.3: Comparison between Neural Networks and traditional machine learning techniques with respect to the amount of data available[59]

our case micro influencer and not micro influencer). We can manipulate the samples considered by varying the C parameter: with small values we use almost every sample in the dataset, while high values are used to consider just the ones close to the margin of the hyperplane. Another important parameter is *gamma* which determine how flexible or rigid is the hyperplane, it works in fact on the kernel. If the gamma value is to large then overfitting could occur.

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right) \quad (11.1)$$

The kernel trick is a mathematical way to learn a nonlinear classification rule which corresponds to a linear classification rule for the transformed data points. So we work in an higher dimension when performing the separation of classes and then we project to a lower dimension to see the transformed function. We create our dataset downloading tweets from the platform so we do not have a test set predefined and for this reason we perform a Stratified Kfold with 5 and 10 folds. We use the Stratified because allows us to maintain micro influencer samples in every round of the fold, in fact it emerges from table that classes are really unbalance with few micro influencer with respect to total users retrieved12.1.

Chapter 12

Experimental Setup

12.1 Dataset: Statistics and Insights

We explore different topics: we start from more environment sensible one because our initial use case was the need of a company working in clean water systems to find micro influencer. The emerging necessity in climate changing and environments risks push us to find micro influencer in topics we consider crucial for the environment. Later we want to not make the predictor too much unbalanced with respect to environment related topics so we explore other sources of users and data: i.e. fintech, womenintech, robotics, etc.

12.2 Validation metrics

balanced accuracy allows us to work with highly imbalanced database. It is the average of recall obtained in each class. This metrics is interesting because it gives us the perception of how the small percentage of users micro influencers over the total users analysed impacts the general result.

accuracy

$$\frac{tp + tn}{tp + tn + fp + fn} \quad (12.1)$$

is useful to detect if something is going wrong, because high values in accuracy and low values on other is a symptome that there are too many badly predicted micro influencer so this must be seen in comparison with the other metrics. In order to delete any doubts it is important to see the confusion metrics that is really clear in our binary case.

recall

$$\frac{tp}{tp + fn} \quad (12.2)$$

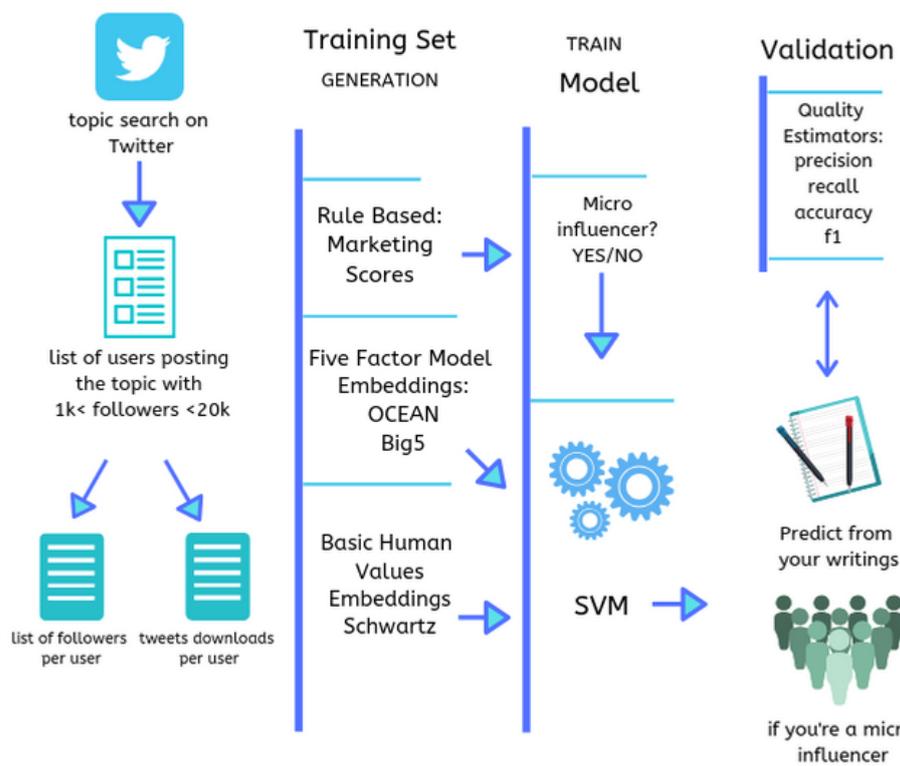


Figure 12.1: Experimental setup from data retrieval to validation

total tweets per user vs. topic

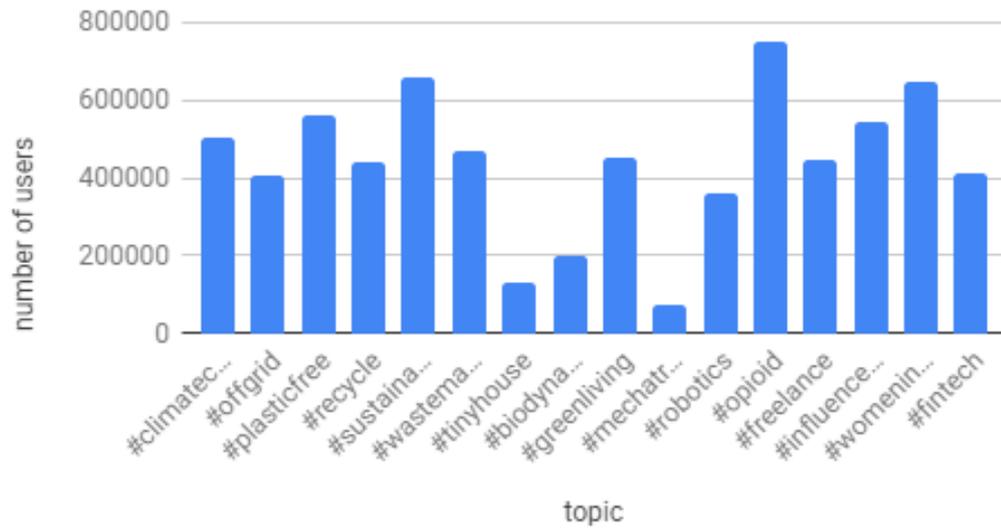


Figure 12.2: Number of tweets downloaded per topic

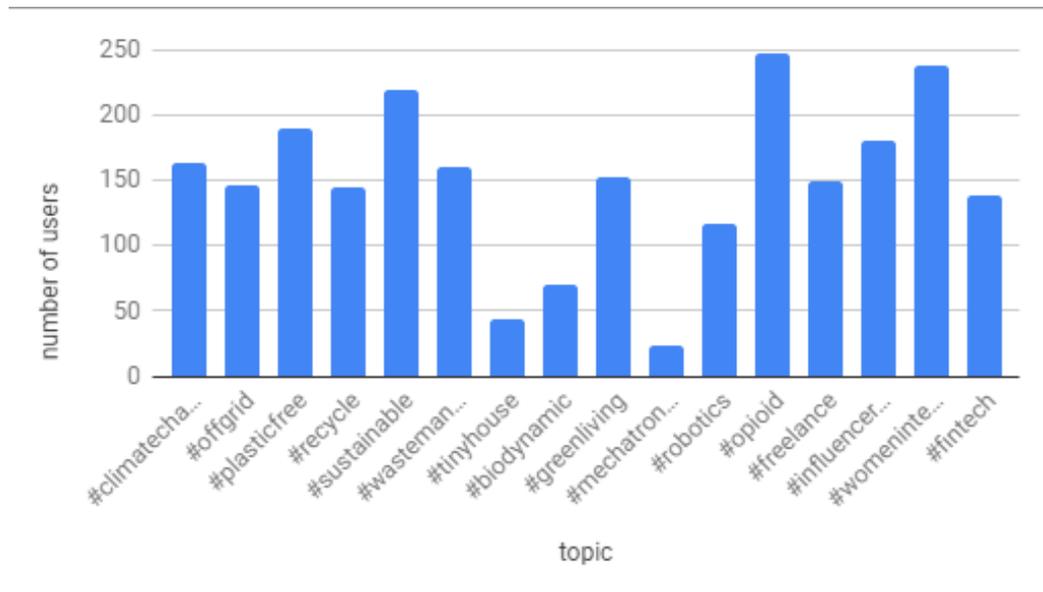


Figure 12.3: Number of potential micro influencer per topic

| Twitter data statistics | | | | | | |
|-------------------------|-----------------|-------------------|------------------------|----------------------|---------------------------|-------------------------|
| Topic searched | Number of users | Micro Influencers | Total tweets per topic | Mean tweets per user | Total followers per topic | Mean followers per user |
| climatechange | 163 | 6 | 504762 | 3096.7 | 997668 | 4280.17 |
| offgrid | 146 | 16 | 407957 | 2794.23 | 589033 | 4034.47 |
| plasticfree | 190 | 12 | 560655 | 2950.82 | 932475 | 4907.76 |
| recycle | 145 | 1 | 439951 | 304.14 | 678766 | 4681.14 |
| sustainable | 219 | 11 | 658854 | 3008.47 | 1010406 | 4613.72 |
| wastemanagement | 161 | 8 | 470334 | 2921.33 | 821268 | 5101.04 |
| tinyhouse | 44 | 0 | 128991 | 2931.61 | 510002 | 5312.52 |
| biodynamic | 70 | 6 | 201010 | 2871.57 | 265217 | 3788.81 |
| greenliving | 153 | 8 | 454312 | 2969.36 | 693278 | 4531.22 |
| mechatronics | 24 | 0 | 71788 | 2991.17 | 86845 | 3618.54 |
| robotics | 117 | 6 | 260898 | 3084.60 | 581328 | 4968.61 |
| opioid | 248 | 23 | 784217 | 3017.01 | 1435238 | 5787.25 |
| freelance | 150 | 3 | 444137 | 2960.91 | 698132 | 4654.21 |
| influencermarketing | 181 | 13 | 544484 | 3010.21 | 921572 | 5091.55 |
| womenintech | 238 | 9 | 644772 | 2709.13 | 989066 | 4155.73 |

Table 12.1: Twitter data Statistics

really highlights the vulnerability in economical terms, because our tool finds micro influencers among a huge amount of non micro influencer users so, if *false negative* is an high number, can lead to few or not found results. We want to keep recall as closer to one as possible and so false negative low.

precision

$$\frac{tp}{tp + fp} \quad (12.3)$$

on the other side is somehow more acceptable because the presence of user that are considered micro influencer and later on, thanks to marketing analysis they are not, can be solved dropping founding to these ones if not effective. But we preserve the choice over an higher numbers of micro influencers.

f1-score

$$2 * \frac{precision * recall}{precision + recall} \quad (12.4)$$

is the armonic mean between recall and precision so can be good to understand the compromise, so not much effort in later cleaning of not micro influencer wrongly predicted and on the other side not leaving out micro influencer not captured. obviously high f1 score, nearly 1, ise the dream of the tool and of the company in the pursuit of maximum optimization.

12.3 Consideration on scoring results

We highlight the values in table 12.2 regarding recall in Support Vector Machine while predicting if a user is a micro influencer or not from just her personality traits. The best performance is obtained in recall of offgrid topic and mechatronics and plasticfree, in these configuration we have the higher number of micro influencers in prportion of the total of user and also their values of embeddedness and recall are high, this allow the output to be solid on the premise so later the supervised learnig can work on more solid bases. But why we are so focused in recall instead of precision? First of all as illustrated in figure 12.4 we cannot have maximum precision and recall at the same time, then on the opposite of medical test we prefer to have less false negative case, because we just have very few positive output and also deleting the ones that are micro influencers not considering them would be deleterious. So we want an high recall that indicates, if near to one, that we have few or none false negative. To achieve this results we force the predict to give more weight on the positive samples through the option class-weight in model definition in scikit learn, in particular we set a proportion of 5 to 1 between micro influencer and not micro influencer. This setting solve cases where we have very little amount of micro influencers catching those

few, but on the other side we lose in precision. The fact that precision is low is not a too bad scenario because we have to understand that we have already performed a lot of filters on users retrieved: filter on number of follower first of all that clean out a huge amount of candidates, then false candidates through granting the acceptance of users who have more followers than following (we do not want bot or spammer), and at the end of the story having much more false positive on the subset is clear that is a little bad training set but at the same time all personality traits are near to the goal so the misunderstanding is not so bad. What we are trying to say is that we are working on really similar in personality traits candidates and companies can leverage on almost suitable candidates, later with marketing analysis on efficiency they can recover investment with low impact if we suggest few micro influencers not effective instead of not giving them the ones really effective because exclude by our tool.

12.4 Results

| Models comparison | | | | | | | | | |
|-------------------|-------|-------------|--------|-----------------|-----------------|------------------|-------|-------|--------|
| Topic searched | SVC p | SVC r | SVC f1 | Rand. For-est p | Rand. For-est r | Rand. For-est f1 | CNN p | CNN r | CNN f1 |
| climatechange | 0.29 | 0.33 | 0.31 | 0.33 | 0.07 | 0.11 | 0.01 | 0.01 | 0.01 |
| offgrid | 0.42 | 1 | 0.58 | 0.49 | 0.44 | 0.50 | 0.63 | 0.31 | 0.42 |
| plasticfree | 0.38 | 0.75 | 0.5 | 0.48 | 0.3 | 0.3 | 0.4 | 0.29 | 0.34 |
| recycle | 0.3 | 0.15 | 0.2 | 0.01 | 0.01 | 0.01 | 0.41 | 0.29 | 0.34 |
| sustainable | 0.23 | 0.47 | 0.3 | 0.48 | 0.11 | 0.14 | 0.6 | 0.27 | 0.375 |
| wastem.. | 0.16 | 0.4 | 0.22 | 0.29 | 0.17 | 0.2 | 0.09 | 0.03 | 0.04 |
| biodynamic | 0.4 | 0.6 | 0.47 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| greenliving | 0.26 | 0.5 | 0.32 | 0.5 | 0.125 | 0.2 | 0.01 | 0.01 | 0.01 |
| mechatronics | 0.52 | 0.83 | 0.62 | 0.58 | 0.61 | 0.60 | 0.54 | 0.64 | 0.58 |
| robotics | 0.24 | 0.48 | 0.31 | 0.66 | 0.20 | 0.30 | 0.33 | 0.16 | 0.21 |
| opioid | 0.21 | 0.65 | 0.32 | 0.51 | 0.31 | 0.36 | 1 | 0.2 | 0.33 |
| freelance | 0.11 | 0.29 | 0.16 | 0.08 | 0.03 | 0.04 | 0.09 | 0.04 | 0.05 |
| influencer... | 0.1 | 0.1 | 0.1 | 0.33 | 0.07 | 0.11 | 0.01 | 0.01 | 0.01 |
| womenintech | 0.21 | 0.55 | 0.30 | 0.27 | 0.05 | 0.07 | 0.307 | 0.09 | 0.14 |
| all preavious | 0.21 | 0.69 | 0.34 | 0.28 | 0.02 | 0.04 | 0.01 | 0.01 | 0.01 |

Table 12.2: Classifier comparison

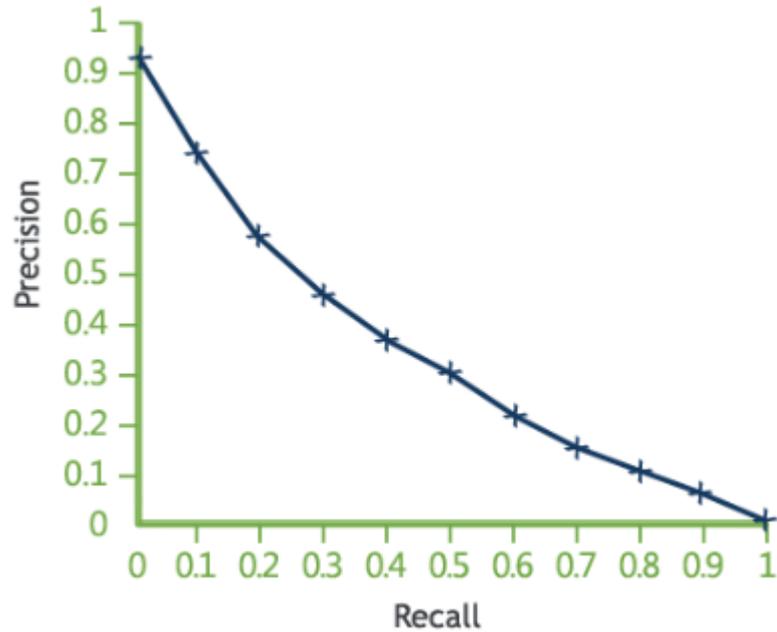


Figure 12.4: Precision and Recall comparison [57]

We can notice the better score is obtained in recall and is a situation we were looking for. High recall indicates that there are few or none false negative, so we do not lose micro influencers that are not detected from our tool. On the other side we know that a low precision indicates many false positive so in a second time the company committing the research over a particular topic will need to detected how many micro influencers found are effective in advertising.

It is important to notice that when very few micro influencers appears, than we can try to relax the threshold to be considered. This reasoning is motivated from the fact that if a topic is not yet mainstream so there will be very few posts authors that write about it, so even small hints of influencer traits must be considered.

12.5 Discussions

It is possible to notice in figure 12.5 that the presence of micro influencers is really low in percentage with respect to all users analyzed. This create a really unbalanced dataset that wouldn't be a problem in the case of a huge one, but in our case doesn't allow the CNN to deep understand and

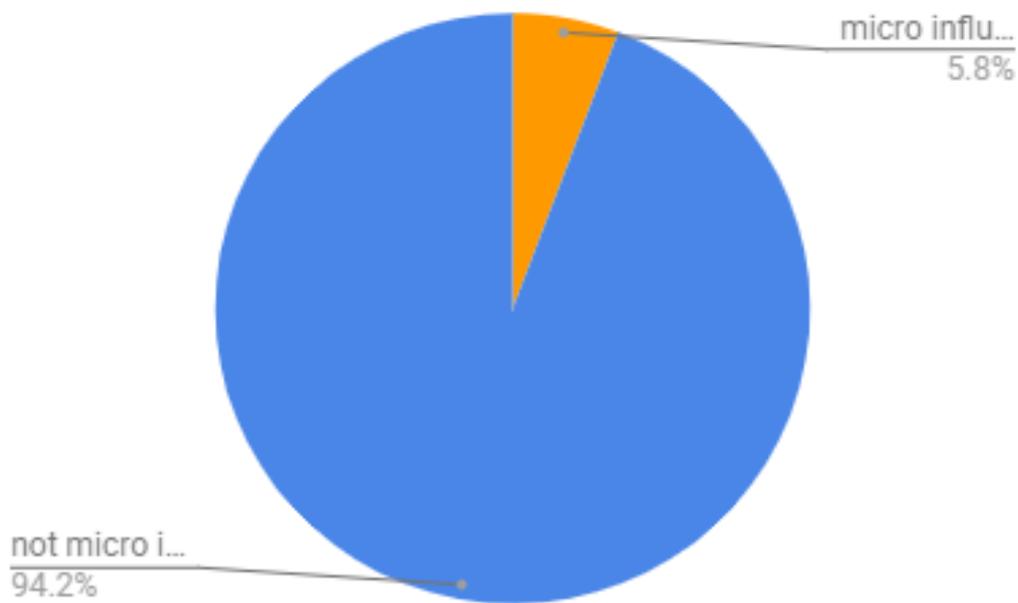


Figure 12.5: Micro Influencer and Not Micro Influencer retrieved

comprehend the definition of micro influencer from features. We can avoid this situation in future expanding the datasets but there will be an embedded problem: micro influencers are hard to find, especially in an early stage. We need to perform analysis over a larger span of time in which the Twitter search command will return different micro influencers over the same topic. Another thing we can see in figure 12.4 is that when we increase precision, recall decreases, so we want to preserve high value of recall with the risk of low precision as described previously.

Part IV

Conclusion and Future Work

In conclusion we demonstrated how an hybrid approach is necessary to converge multiple disciplines to reach the common goal of general improvements in all subjects involved. The supervised learning model of SVM is at the moment able to determine if a user is or not a potential micro influencer of a specific topic just learning from user's writing and extracting his personality traits. We answered both our research questions: first it is possible, in fact, to define if a user is a micro influencer of a topic given by a company looking for brand ambassadors and second we correlated psychological traits with the output given starting from marketing parameters and later derive similar results just starting from Big5 Values and Basic Human Values. Large margin of improvements are possible in the field of tool generalization such as find values of a generic micro influencer instead of a specific topic one. To achieve this last step it is necessary to expand the dataset retrieve and wait for a better definition commonly accepted of micro influencer from economics area and would be challenging work in strict relationship with a group of sociologist and psychologists to define an psychological identity behind the figure in consideration. Further improvements can be achieved in Big5 and Basic Human Values retrieval from text adopting a gold standard with which we can compare results given by our tools and the ones given by a psychological survey. This tool is at its first steps in the way of companies practical utilization but it has all the possibilities of becoming a powerful help in recruiting and nonetheless many others applications. One of them could be using over mobile application and social platform to better detect micro influencer in a passive way and give them direct contact with companies which want to hire them. As shown in the section where talking about Idiap work on YouTube dataset would be extremely powerful to exploit also video audio details of a user, retrieving these information from vlog on YouTube or Instagram images and stories. We would create an inversion in trends in a way that Artificial Intelligence can propose to the user something new. Here the concept of serendipity emerges: nowadays AI reacts to human suggestions and behaviours, without trying to elaborate conditions in which human can evolve escaping his comfort zone. We would to embed in future something similar in our work to generate suggestions. One more improvement would be focus more on the timestamp of data retrieval to determine the user evolution over time, this is fundamental in the passage from adolescence to adult life where personality traits can change drastically and interest evolve and transform the user vision of the world and of the surrounding environment. Geographical position is ignored in our work due to the ambition of Shalom Schwartz [2] of define a worldwide common model across cultures, but we have seen that an intra-community inspection of interests and language can certainly improve the output of the tool. We also

worked on another intuition: instead of just focusing on micro influencer side why can't we look also at her audience? During CIKM conference in 2018 at Lingotto, Turin many authors tried to understand [\[45\]](#) user groups behaviour in social media. This situation can be treated in future work to understand if people attracted by certain topics have common personality traits or if they are just impressionable as in the case of hypnosis where certain patients react well to cure and are not at all affected.

Bibliography

- [1] Anger,I.,Kittl,C., Measuring influence on Twitter. In Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies (i-KNOW '11), ACM, Article 31, 4 pages (2011)
- [2] Schwartz, S. H., An Overview of the Schwartz Theory of Basic Values. Online Readings in Psychology and Culture, 2(1), (2012).
- [3] McCrae, R. R., Costa, P. T., Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52(1), 81-90, (1987).
- [4] Barcelo J., National Personality Traits and Regime Type: A Cross-Cultural PSychoLOGY, 48(2), pp 195-216, (2017).
- [5] Matsumoto D., Hwang H. S., Culture and Emotion: The Integration of Biological and Cultural Contributions, *Journal of Cross-Cultural Psychology*, 43(1), 91-118, (2012).
- [6] Datler G., Jagodzinski W., Schmidt P., Two theories on the test bench: Internal and external validity of the theories of Ronald Inglehart and Shalom Schwartz, *Social Science Research*, Vol 42, Iussue 3, pp 906-925, (2013).
- [7] Roccas, Sonia and Sagiv, Lilach and Schwartz, Shalom and Knafo-Noam, Ariel. The Big Five Personality Factors and Personal Values. *Personality and Social Psychology Bulletin*. 28. 789-801. (2002).
- [8] J. Dollinger, Stephen and Leong, Frederick and K. Ulicni, Shawna. On Traits and Values: With Special Reference to Openness to Experience.*Journal of Research in Personality*. 30. (1996).
- [9] Y.R. Tausczick and J. W. Pennebaker, The psychological meaning of Words: LIWC and Computerizes Text Analysis Methods, *Journal of Language and Social Psychology*, 29(1), pp.24-54, (2010).

- [10] Chen, Jilin and Hsieh, Gary and Mahmud, Jalal and Nichols, Jeffrey. . Understanding individuals' personal values from social media word use. Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW. 405-414. (2014).
- [11] McAdams, Pals JL., A new Big Five: fundamental principles for an integrative science of personality. *Am Psychol.*,61(3):204-17, (2006).
- [12] Gudonis, Lauren C., "The interaction between personality traits and contextual disadvantage on criminal behaviour: longitudinal study of high risk-females". University of Kentucky Doctoral Dissertations. 742. (2009).
- [13] Reganti, Aishwarya N. et al. Semantic Interpretation of Social Network Communities (AAAI Student Poster Additional). (2016).
- [14] Ackermann, K. and Ackermann, M., The Big Five in Context: Personality, Diversity and Attitudes toward Equal Opportunities for Immigrants in Switzerland. *Swiss Polit Sci Rev*, 21: 396-4180, (2015).
- [15] Boyle, Gregory J. Current research in personality traits and individual differences. *Humanities and Social Sciences papers.*(2010).
- [16] Allport, G.W. *Personality: A psychological interpretation.* New York: Holt. (1937).
- [17] Chiu C. Y., Kim Y. H. and Wan W.W.N., *Personality: Cross-cultural perspectives. Handbook of personality theory assessment; vol. 1. Personality theories and models.* pp 124-144. Los Angeles. (2018).
- [18] McCrae R. R. and Costa P.T., Empirical and theoretical status of the Five Factor Model of personality traits. *Handbook of personality theory assessment; vol. 1. Personality theories and models.* pp 273-294. Los Angeles. (2008).
- [19] Simon J., Perez-Testor, et al., The Portrait Values Questionnaire: A Bibliographic and bibliometric review of the instrument. *Revista de Psicologia, Ciencias de l'Educatio y de l'Esport.* 35(1), pp39-50, (2017).
- [20] Costa, Paul and R. McCrae, Robert. The revised NEO personality inventory (NEO-PI-R). *The SAGE Handbook of Personality Theory and Assessment.* 2. 179-198.(2008).

- [21] A. Mandelbaum, Adi Shalev, Word Embeddings and Their Use In Sentence Classification Tasks, arXiv e-prints, 13 pages. Hebrew University of Jerusalem. (2016).
- [22] Zook M, Barocas S, boyd d, Crawford K, Keller E, Gangadharan SP, et al. Ten simple rules for responsible big data research. PLoS Comput Biol 13(3). (2017).
- [23] E. Bakshy, J. M. Hofman, W. A. Mason, D. J. Watts, Everyone’s an influencer: quantifying influence on twitter, Proceedings of the fourth ACM international conference on Web search and data mining, Hong Kong, China, P 65-74, (2011).
- [24] Bigonha, et al. Sentiment-based influence detection on Twitter. Journal of the Brazilian Computer Society. (2011).
- [25] Kumar U., et al. Inducing Personalities and Values from Language Use in Social Network Communities, Springer, vol 20, iussue 6, pp 1219-1240, (2018).
- [26] Robert West, Hristo S. Paskov, Jure Leskovec, Christopher Potts, Exploiting Social Network Structure for Person-to-Person Sentiment Analysis, arXiv e-prints, Cornell University, 2014.
- [27] T. Weia, Y. Luc et al., A semantic approach for text clustering using WordNet and lexical chains, Volume 42, Issue 4, Pages 2264-2275, (2015).
- [28] Wu Youyou, Michal Kosinski, and David Stillwell. *Computer-based personality judgments are more accurate than those made by humans*. Proceedings of the National Academy of Sciences, 112(4):1036-1040, 2015.
- [29] Barcelò Joan. *Cultural involvement: Geert Hofstede’s cultural factors [National Personality Traits and Regime Type: A Cross-National Study of 47 Countries]*. Journal of Cross-Cultural Psychology, 48(2):195-216, 2017.
- [30] Mackenzie, Emma. Celebrity endorsement. BandT, No. 2811, 86-88, 90-91, (2015)
- [31] Biel, Joan-Isaac and Gatica-Perez, Daniel. Vlogcast Yourself: Nonverbal Behavior and Attention in Social Media. 10.1145/1891903.1891964. (2010).
- [32] Joan-Isaac Biel, Oya Aran, Daniel Gatica-Perez, You Are Known by How You Vlog: Personality Impressions and Nonverbal Behavior in

YouTube.AAAI Publications, Fifth International AAAI Conference on Weblogs and Social Media. (2011).

- [33] Biel, Joan-Isaac and Gatica-Perez, Daniel, The youtube lens: Crowd-sourced personality impressions and audiovisual analysis of vlogs, *Multimedia, IEEE Transactions on*, volume 15, number 1, pp 41-55, IEEE, (2013).
- [34] Peabody, Golberg, Some determinants of factor structures from personality traits descriptors. *Journal of Personality and Social Psychology*. 57(3): 552-567. 1989.
- [35] Jure Leskovec, Lada Adamic, and Bernardo Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 1(1), May 2007.
- [36] Pennebaker, J., Niederhoffer, K., Mehl, M. Psychological aspects of natural language use: Our words, ourselves. *Annual Review of Psychology*. 54:547-577, (2003).
- [37] Pennebaker, J., King, L.A. Linguistic Styles: Language Use as an Individual Difference. *Personality and Social Psychology*, 1999, 77(6):1296-1312.
- [38] Myers, I. B. *The Myers-Briggs Type Indicator: Manual* (1962). Palo Alto, CA, US: Consulting Psychologists Press. (1962).
- [39] Lindeman, Marjaana and Verkasalo, Markku. Measuring Values With the Short Schwartz's Value Survey. *Journal of personality assessment*. 85. (2005).
- [40] Max Weisbuch, Zorana Ivcevic, On being liked on the web and in the "real world": Consistency in first impressions across personal webpages and spontaneous behavior, *Journal of Experimental Social Psychology*, Volume 45, Issue 3, May 2009, Pages 573-576
- [41] Schwartz, Shalom H, and Romie F Littrell. "Draft User's Manual. Proper Use of the Schwartz Value Survey." Unpublished manuscript, Auckland, New Zealand (2007).
- [42] Fast, Ethan and Chen, Binbin and Bernstein, Michael, Empath: Understanding Topic Signals in Large-Scale Text, *arXiv e-prints*,(2016).
- [43] Carducci, Giulio and Rizzo, Giuseppe and Monti, Diego and Palumbo, Enrico and Morisio, Maurizio, *TwitPersonality: Computing Personality*

- Traits from Tweets Using Word Embeddings and Supervised Learning, Information,9,(2018)
- [44] Landauer, T.K., Dumais, S.T. A solution to Plato's problem: The Latent Semantic Analysis theory of the 627 acquisition, induction, and representation of knowledge Psychological Review, 1997, 104:211-240.
- [45] Sihem Amer-Yahia, Behrooz Omidvar-Tehrani, Joao Comba, Viviane Moreira, Fabian Colque Zegarra. Exploration of User Groups in VEXUS. arxiv. International Conference on Data Engineering (ICDE) 2018.
- [46] Youyou, Wu, Kosinski, Michal, Stillwell, David, Computer-based personality judgments are more accurate than those made by humans, Proceedings of the National Academy of Sciences, Proc Natl Acad Sci USA, 112, (2015).
- [47] Fabio Celli, Bruno Lepri, Elia Bruni, Automatic Personality interaction style recognition from facebook profile pictures. In proceedings of the ACM International Conference on Multimedia, 2014.
- [48] Fabio Celli and Luca Rossi, The role of emotional stability in Twitter conversations, in proceeding od the workshop on semantic analysis in social media, Association for computational linguistics. 2014.
- [49] Fabio Celli, Elia Bruni, Bruno Lepri, Automatic Personality and Interaction Style Recognition from Facebook Profile Pictures, Proceedings of the 22nd ACM international conference on Multimedia, Orlando, Florida, USA, (2014).
- [50] Albert Einstein. *Zur Elektrodynamik bewegter Korper*. [*On the electrodynamics of moving bodies*]. Annalen der Physik,322(10):891-921,1905.
- [51] Knuth: Computers and Typesetting,
<http://www-cs-faculty.stanford.edu/~uno/abcde.html>
- [52] Easley,D.,Kleinber,J.,Networks, Crowds, and Markets:Reasoning about a Highly Connected World, 819. Cambridge University Press, Cornell University (2010), 59-69
- [53] B. R.Little , Free Traits, Personal Projects and Idio-Tapes: Three Tiers for Personality Psychology, Psychological Inquiry, vol 7, num 4, 340-344,Routledge, (1996).

- [54] Richard W. Robins, R. Chris Fraley, Robert F. Krueger, Handbook of Research Methods in Personality Psychology, 719. Guilford Press, New York, (2009)
- [55] Francis Galton, Measurement of Character, 1884
- [56] Carl Jung, Psychological Types, 634. Routledge. U.K. (1921).
- [57] <https://www.searchtechnologies.com/precision-recall>
- [58] <https://www.16personalities.com/enfj-personality>
- [59] Bahnsen Alejandro, Correa. Building ai applications using deep learning. 2016. URL <https://blog.easysol.net/wp-content/uploads/2017/06/image1.png>
- [60] Cameron Marlow, Lee Byron, Tom Lento, and Itamar Rosenn. Maintained relationships on Facebook, 2009. On-line at <http://overstated.net/2009/03/09/maintainedrelationships-on-facebook>.
- [61] McKenzie Andre, Kashef Ijaz, Jon D. Tillinghast, Valdis E. Krebs, Lois A. Diem, Beverly Metchock, Theresa Crisp, and Peter D. McElroy. Transmission network analysis to complement routine tuberculosis contact investigations. American Journal of Public Health, 97(3):470477, 2007.