ESCOLA
SUPERIOR
DE TECNOLOGIA
E GESTÃO
POLITÉCNICO
DO PORTO

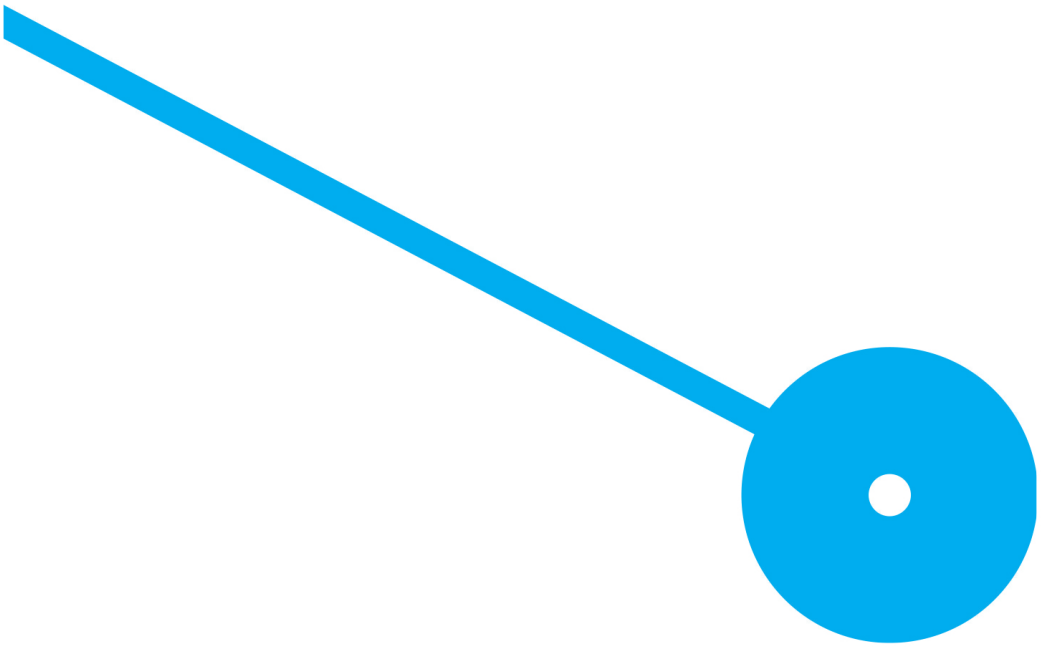
M

MESTRADO
Engenharia Informática

Individuals Recognition and Simulation Based on Multiple Data Sources

Ricardo Manuel Pacheco Barbosa

11/**2017**



Dedicatory

Este trabalho não poderia estar concluído sem antes expressar os contributos que não podem e nem devem deixar de ser realçados. Por essa razão, desejo expressar os meus sinceros agradecimentos:

Ao Professor Doutor Ricardo Jorge da Silva Santos, o meu orientador, pela competência científica, acompanhamento ao longo de todo o projeto, pela disponibilidade, pela ajuda prestada, pelas críticas construtivas, e pelo conhecimento transmitido durante os anos na qualidade de docente e amigo.

À minha namorada Jéssica de Sousa, por todo o apoio permanente, expresso ou silencioso, por ouvir as minhas opiniões, por discutir comigo todas as minhas ideias, por estar presente nos momentos mais difíceis, e por toda a compreensão e encorajamento.

Ao meu pai Ângelo Barbosa e à minha mãe Maria Jacinta Pacheco, por acreditarem em mim, pelo apoio que me dão e pelo esforço financeiro que realizaram durante este meu percurso académico.

Aos meus avós que, apesar de uma vida inteira de trabalho físico árduo, sempre valorizaram, incentivaram, e privilegiaram a educação e a formação académica de todos os seus descendentes.

A todos eles dedico este trabalho.

Abstract

The increase of popularity and usage of Internet platforms, such as online social networks, has resulted in a growth of volume of data and, most specifically, personal data. The presence of ambient intelligence and smart environments on our daily life is associated with a constant need of data that is usually gathered trough a vast collection of physical sensors. This urge for data is the motivation for the first model that is proposed on this work. This model consists on a virtual social sensor that extracts data from online social networks, to produce knowledge and wisdom to smart environments, and identifies personality characteristics to predict behaviour and prepare appropriate interactions when facing different types of events. Online social networks are also platforms that allow users to construct their identities via self-representation, which includes their interests. The second model proposal of this work is a multidimensional interests network that intends to aid organisations and smart environments with insights about an individual (or a group of individuals), measure the impact of online content, interests, and understand patterns of usage.

Keywords: Online Social Networks, Smart Environments, Behaviour, Personality Identification, Virtual Social Sensor, Multidimensional Interests Network

Resumo

O crescimento e o aumento de popularidade da Internet, originou o aparecimento de serviços, tais como as redes sociais, que se tornaram rapidamente populares devido à sua capacidade de criação de um perfil virtual que pode, ou não, conter semelhanças com a vida real do detentor desse perfil. Esta criação, e constante construção de uma identidade virtual, está diretamente associada com características psicológicas designadas por "need to belong" e "need for self presentation".

Esta forte presença de utilizadores e consequente utilização destes serviços, resulta numa presença constante de dados pessoais que, para além de características demográficas, contém informação relativa aos interesses e às preferências dos seus utilizadores, bem como características que permitem a identificação de traços de personalidade. A personalidade, tal como outras características pessoais, é um identificador único de cada um de nós e está diretamente relacionada com os diferentes comportamentos que expressamos em determinados contextos. O estudo da personalidade está relacionado com a área da psicologia e, normalmente, a sua identificação é realizada através do preenchimento de questionários que podem ser morosos ou tediosos. Alguns estudos nesta área conseguiram identificar a presença de traços de personalidade em texto escrito, formal ou informal, que elimina a necessidade do preenchimento de questionários para a obtenção de uma referência da personalidade de um indivíduo.

As mudanças do mercado em volta dos interesses dos consumidores, levou a uma mudança estratégica por parte das organizações que, acompanhando a tendência dos seus clientes, decidiram adotar as plataformas sociais virtuais como plataformas para divulgação de conteúdo, e de aproximação aos seus clientes. Esta adoção representa também uma enorme quantidade de dados que está imediatamente disponível para a organização, ao contrário do que antes era realidade. Esta súbita quantidade de dados, e diferentes níveis de complexidade associada aos mesmos, resulta numa incapacidade das organizações de obtenção de informação ou conhecimento a partir dos mesmos. Apesar de poderem conseguir obter alguma informação, existe ainda um nível mais vasto de informação e conhecimento que está embebido no conteúdo e nas interações que são realizadas.

Este vasto volume de dados cada vez mais presente nos dias de hoje, originou também o crescimento de áreas que beneficiam deste grande volume de dados. O aumento do volume de estudo em ambientes inteligentes tem sido mais incidente nos últimos anos como resultado natural do aumento do volume de dados, e da capacidade de recolha dos mesmos. Tal como outras áreas, os ambientes inteligentes também beneficiam da obtenção de informação e conhecimento sobre as características pessoais e os interesses dos seus habitantes, de forma a poder adaptar o ambiente de acordo com as necessidades e preferências dos mesmos. A presença, nas plataformas sociais presentes na Internet, de dados que podem identificar características pessoais e preferências, tornam estas plataformas uma fonte de dados promissora para o melhoramento e o aumento das capacidades presentes num ambiente inteligente.

Como resultado desta necessidade de obter informação pessoal e um nível de conhecimento maior acerca de um conjunto de indivíduos específico, esta dissertação propõe dois modelos: um sensor virtual baseado nas plataformas sociais, e uma rede multidimensional de interesses.

O primeiro modelo é focado no conteúdo e nas interações efetuadas nas plataformas sociais online, para identificar características associadas aos traços de personalidade dos utilizadores, processar e compreender o conteúdo presente na forma de texto escrito, extrair e detetar emoção e sentimento presente nesse conteúdo, bem como uma constante monitorização, deteção e identificação de interações. Este modelo apoia-se nos estudos focados na extração de traços de personalidade através de texto escrito, para adotar as plataformas sociais online como uma fonte de dados para a obtenção de informação pessoal e características associadas à personalidade, que, influenciam também o comportamento que é demonstrado. Esta solução procura fornecer um nível detalhado de informação pessoal para as organizações e permitir a identificação dos interesses dos seus clientes, obtenção da opinião deles em relação a um tópico ou a um produto, identificação de características demográficas, identificação de potenciais clientes, ou até mesmo a medição do impacto dos seus conteúdos ou das estratégias de marketing. De uma forma semelhante, este modelo fornece aos ambientes inteligentes novos dados que contém informação relativa às preferências e interesses dos seus habitantes. Este sensor, que apesar de ser virtual, cumpre todos os requisitos necessários na definição de um sensor no contexto de um ambiente inteligente, e é uma forma não obstrutiva de recolha de informação sem a necessidade de pedir diretamente (dado que os utilizadores disponibilizam esses dados de forma voluntária).

O segundo modelo presente neste trabalho é inspirado no estudo das redes complexas e na representação de cenários do extraidos do mundo real. O estudo das redes complexas tem sido utilizado nos contextos sociais, tecnológicos, industriais, ou biológicos para representar interações entre pessoas, redes de transporte presentes numa cidade, rede de serviços que fazem parte de uma habitação, ou até mesmo representar estruturas moleculares. Face à necessidade de identificar os interesses de indivíduos, este modelo apresenta uma abordagem multidimensional que, ao contrário dos cenários atuais em relação aos interesses individuais, classifica os interesses em diferentes níveis e com diferentes valores de impacto. Esta representação multidimensional aproxima-se dos cenários do mundo real, em que os interesses possuem diferentes tipos de intensidade ou importância, para fornecer às organizações ou aos ambientes inteligentes, uma nova perspetiva e abordagem relativamente aos interesses dos seus clientes/habitantes. A multidimensionalidade associada a este modelo, bem como a representação em camadas, resulta numa possibilidade de comparação de diferentes interesses entre vários indivíduos, respondendo assim de forma direta à necessidade dos ambientes inteligentes em considerar os interesses de vários indivíduos que estejam presentes no mesmo ambiente, em função de considerarem apenas as necessidades e interesses de um único individuo.

Contents

List of Figures

List of Tables

List of Source Code

List of Abbreviations

AI	**A**rtificial **I**ntelligence
AmI	**Am**bient **I**ntelligence
ISTAG	**I**nformation **S**ociety **T**echnologies **A**dvisory **G**roup
NLP	**N**atural **L**anguage **P**rocessing
NLTK	**N**atural **L**anguage **T**ool **K**it
OSN	**O**nline **S**ocial **N**etwork
POS	**P**art **O**f **S**peech
SmE	**Sm**art **E**nvironments
SVM	**S**upport **V**ector **M**achine

Chapter 1

Introduction

In the evolutionary step, humans clearly distinguish themselves from other mammals with characteristics such as opposable thumbs, or a logical mind and thoughts that have developed trough the ages. Those clear distinctions are also present when we discuss the differences among humans. From characteristics that includes fingerprints, molecular composition, or even personality, humans possess unique individual aspects, and there is a constant work regarding the classification and identification of those unique characteristics. Among all other characteristics that distinguish humans, personality is an uniquely identifier that also impacts and moderates human behaviour.

Personality is directly correlated with the psychology study domain, and its classification is usually achieved by filling out questionnaires, however, some studies found a correlation between written text and personality traits which opens the possibilities and the applicability opportunities for the inclusion of personality in other domains, avoiding the necessity of filling a questionnaire (which can be a tedious procedure).

In a world marked by revolutions and eras, it is possible to affirm that we live in the information era, where the amount of data that is produced today is far superior when compared to the past. This large increase of data available today allowed some areas to boost their applicability and their potential, or support the study and the emergence of new areas. Some areas date the 60's, like the artificial intelligence, but only now their potential is becoming more clear due to the presence, and the constant generation, of a large amount of data that supports and allows the application of solutions that were not possible before. Ambient Intelligence and Smart Environments are impulsed by ubiquitous computing and take advantage of the ease of collecting data from numerous devices in order to produce tasks such as optimisation of energy consumption, recognition of human activity and preferences, aid the elderly or persons with health problems, or increase the lifestyle of blind people (as examples). This can be extended to numerous scenarios that are not restricted by the boundaries of a house (or smart house), and it can be applied in a "micro" scale or in a large scale like the concepts of smart cities.

With the emergence and the constant requirement for more adaptable and customised solutions, became the necessity of gathering even more data and, more specifically, more personal data that can provide insights about preferences or even patterns of behaviour. The growing of popularity and usage of the Internet, and its platforms, is responsible for the generation of enormous amounts of data day after day. One of the reasons behind this large amount of data generation is related to the popularity of a specific segment of Internet platforms, the online social networks (and other type of social platforms). With the most popular online social networks surpassing a milliard (one thousand million) of active users, they

have a strong impact on the daily life of people that adopt these platforms as a way to construct a virtual identity and present themselves to the world. This usage is also characterised by the presence of personal data, in different degrees of accuracy, that is expressed deliberately or encoded on the content of the interactions that online users perform.

This virtual profiles that are constructed and updated daily from a large amount of people in the world, often represents a virtual mirror to their characteristics in the real world (even if the virtual profile is only a characterisation of the image they would like to pass to others). With this high presence on the daily lives of their customers, even organisations have adopted online social networks as a channel of communication, marketing platform, or even a generalised web image (that replaces the traditional organisational website).

The change in the market around the interests and needs of the consumers, also lead organisations to adopt new methods that allows them to obtain more insights about their consumers and even correctly identify their preferences or needs. The adoption of online social platforms provided organisations with a large amount of data related to their customers, but most organisations are not generating the information and knowledge they need to truly fulfil the needs and interests of their customers. In the previous context of smart environments, there is also a constant need for the environment knowing and adapting around the interests and needs of its inhabitants.

This work, titled "Individuals Recognition and Simulation Based on Multiple Data Sources", is focused on four fields of study (personality characteristics and influence on behaviour; the growth of the internet and the motivations related to the usage of its platforms; ambient intelligence and smart environments; and the study on complex network either by a monodimensional or multidimensional perspective) and proposes two model definitions: a virtual online social network based sensor; and a multidimensional interests network.

The first proposed model present in this work, intends to fulfil the same characteristics and requirements that define a sensor in a smart environment, but is not a physical device and is only focused on data present online, more specifically on online social platforms. Based on the results obtained by other works on personality classification by linguistic aspects present on written text, and the presence of large amounts of data on these online platforms resulted in a logical association between the two subjects. This model intends to fulfil the organisational requirements, and the personal requirements associated with smart environments, by profiling users based on their virtual identities, process natural language that is associated with the written content present on those online platforms, detect emotion and sentiment expressed by written content, and monitor and analyse network interactions which contain interests indicators.

The multidimensional interests network model is based on the study of complex networks and the representation of real world scenarios. From the representation of real world social networks (that are the basis behind the online social networks), to informational networks, power grid networks, transportation routes, or even biological characteristics, almost every type of real world scenario can be represented by a network. This model is part of the network interactions module that is present on the virtual social sensor model, but can be applied on an independent form. The intention of this model is to follow the tendencies of the representation of real world scenarios into complex networks, by representing interests on a multidimensional perspective. These interests can be gathered from the online interactions produced

by an individual and can contain degrees of impact and be adjusted and constructed to fulfil the needs of a particular context (either by an organisational or personal perspective).

1.1 Structure

After the introduction present on this chapter, this dissertation is structured as follows. Chapter two is focused on the study on human personality, from the definition, to the direct influence on behaviour, as well as the explanation of a personality classifier known as the Big Five classifier. Chapter three introduces the definition of Ambient Intelligence and Smart Environments, and the characteristics required in order to have a truly Smart Environment; the chapter ends with a detailed overview focused on the sensors that are present on a Smart Environment. Chapter four includes an overview on the growth of the Internet and its platforms; this chapter makes allusion to the reasons that drive people to use online social platforms and the motivations associated to that behaviour. Chapter five contains a review on the study of complex networks, more specifically, the monodimensional and multidimensional aspects of complex networks and some variations like the multilayer network perspective proposed by other authors. Chapter six is dedicated to the definition of the models that are proposed on this dissertation; is divided into two main parts being the first one dedicated to the definition of the virtual social sensor (and its modules), and the second part being dedicated to the definition of the multidimensional model. Chapter seven contains some theoretical scenarios of application of the models proposed by this work either by a organisational or a personal perspective. This dissertation ends, in chapter eight, with a synthesis overview on the proposed work, with the enumerations of future work, and the scientific contributions that resulted from the creation process of this dissertation.

Chapter 2

Human Cognition

This chapter aims to provide an overall vision around the topic of personality and its correlation to human behaviour. In a first instance, this chapter contains the various definitions of personality trough the times which reveals the controversy around this subject, as well as an overall description of personality that is largely accepted trough the psychology community. The five major theories of personality are also described on this chapter in a lecture overview, and there is a more detailed overview on the Big Five Traits personality model, that is one of the models used to classify personality traits of an individual. This chapter ends with a correlation between personality traits and human behaviour, as well as an literature overview of using written language to extract those traits.

2.1 Personality

Since the remote times of ancient Greece, and the time of Aristotle, that is acknowledged the existence of various personality types and their connection and relation with different patters of human behaviour, and today the definition of personality is still ambiguous. (Poria et al. 2013)

How it is possible to know if a person is dominant or submissive, if approves or disapproves some topic, if likes or dislikes something? These traits or characteristics are not possible to observe physically since they are not part of a person physical characteristics, neither we have direct access to the person's thoughts and feelings. Personality traits and attitudes are latent, hypothetical characteristics that can only be inferred from external observable clues (behaviour, verbal or non verbal communication, and context where the behaviour occurs) (Ajzen 2005). The ability to predict personality has implications in many areas, and existing research has shown connections between personality traits and success in both professional and personal relationships (Jennifer Golbeck et al. 2011).

2.1.1 What Is Personality?

The chronological overview of personality definition, present on the work of Udo-Imeh et al. (2015), explains the controversy around this subject trough the years. From being "something which is difficult to explain in one sentence. It is very vast and dynamic...", to being considered by other authors as a "collection of individual characteristics that make a person unique, and which control an individuals responses and relationship with the external environment", or even "a person's consistent responses to

recurring situations", it is clear that most of those definitions have some incidence and keywords in common.

Despite not having an official definition, most psychologists consider personality as "a dynamic organisation, inside the person, of psychophysical systems that create the characteristics patterns of behaviour, thoughts and feelings of an individual" (Sulaiman, Rambli and Halim 2011). Personality is considered a key component to identify a profile, and an uniquely identifier for each one of us which affects a lot of aspects of human behaviour, mental processes, and affective reactions (Markovikj et al. 2013). Personality is an important factor in social interactions, some people are more talkative while others can be more shy, the same way that some can be more calm while others can be more insecure (Ghavami et al. 2015).

In essence, personal tendencies are shaped further through social interactions where individuals in a social network act similarly, sometimes referred to as normative (or normal) behaviour. The way we talk, act, and write is different from person to person. A simple posture can express some insights about someone, and even when the content of a message is the same, individuals express themselves verbally with their own distinctive styles (Pennebaker and King 1999). However behaviour is not simply a function of personality traits, but personality is an important trait that moderates people's behaviour and interactions with others.

2.1.2 Theories On Personality

The major theories on personality can be grouped into five major theories (Udo-Imeh et al. 2015):

- Pshychodynamic Theory: This theory is founded on the idea that human personality is developed primarily as a result of the interaction and unconscious forces within the individual. The theory assumes that human behaviour is unconsciously driven, that different parts of the unconscious mind are in perpetual conflict, and that our behaviours can be traceable to our childhood experiences. Pshychodynamic theory include the work of Sigmund Freud (psychoanalytic theory) and those of his followers (known as Neo-Freudian Theories). The authors evidence the criticism around this theory, that is focused on the fact that hypotheses generated are not scientifically testable, and the theory is based on the development of personality only on the first five years of life (ignoring the impact of later life experiences);
- Traits Theory: This theory suggests that personality is made up of a set of quantitative measurable characteristics or units, known as traits. Traits are defined as a relative stable tendency to behave in a particular way across a variety of situations and are described as a dimension which people differ from one to another. Each personality consists on a unique combination of traits, and people with a given combination of traits can be expected to behave consistently across situations over time. In the same way, people with similar combination of personality traits can be expected to act similarly in certain situations. The authors consider that this approach is only focused on the description of traits and does not account its development;
- Behavioural Theory: This theory contends that personality is the outcome of the interaction between individual factors and environmental influences. Unlike theories like the psychodynamic theory

and the traits theory, this theory prioritises the observable and measurable external events in function of the inner mental states. To the behaviourists, a person's mind is in a blank state at birth, and personality is then acquired trough conditions and shaped by the reinforcement in the form of rewards and punishment. According to the authors, this approach is too deterministic and assumes that humans do not have or do not exercise freewill;

- Humanistic Theory: This approach to personality states that humans are largely responsible for their actions and have an innate need for personal development and fulfilment in life. It is focused on the subjective and holistic view of human existence and pays particular attention the the issues of creativity, freewill, as well as human potentials. The humanistic theory rejected the deterministic perspectives of the pshycodynamic theory and the behavioural theory, and states that a person's behaviour and choices are determinate by himself and not by fate. The criticism from the authors is related to the vaguely defined key concepts of this theory;
- Socio-congnitive Theory: This theory fuses the cognitive approach to personality with the social learning perspective, due to this fact it is often referred as 'socio-cognitive theory'. The social part of this theory is an extension and modification of the behavioural theory. The cognitive part is focused on the differences in personality as well as the different ways that people process information. The authors consider this theory as subjective and vague for scientific study;

The proposed solution that is present on this work is focused on the traits theory and on the quantitative measurable characteristics that compose it, as well as existing works that are based and focused on those same characteristics.

2.1.2.1 Big Five Factor Model

Despise the several well studied personality models that have been proposed through the years, the Big Five Factor Model, introduced by Norman in 1963 (Ghavami et al. 2015) and matured by Goldberg (Goldberg 2006), was established as the most popular one and is currently the most widespread and generally accepted model of personality (Bachrach et al. 2012; Jennifer Golbeck et al. 2011; Markovikj et al. 2013). Accordingly to the personality theories (Udo-Imeh et al. 2015) this model of personality is based on the concepts of the Traits Theory.

A personality trait is defined as a characteristic of an individual that exerts pervasive influence on a board range of trait-relevant responses, and the Big Five Model (or Five Factor Model) classifies and divides personality according to five different traits (as the name suggests). This model is also often represented under the acronyms OCEAN and CANOE (that are a result of the first letter of each trait).

The five dimensions can be described as the following:

- Openness to Experience: curious, intelligent, imaginative. People that have a high value on this trait tend to be artistic and sophisticated in taste and appreciate diverse views, ideas, and experiences. It can be described as being insightful vs unimaginative;
- Conscientiousness: responsible, organised, persevering. Conscientious individuals are extremely reliable and tend to be high achievers, hard workers, and planners. Organised people have a high value on this trait, while careless people have a low value;

- Extroversion: outgoing, assertive. Extroverts draw inspiration from social situations and are often classified as being friendly and even energetic. High scorers are really sociable, while the ones that possess a low value of extroversion trait are defined as being more shy or quiet;
- Agreeableness: cooperative, helpful, affection. People with a high agreeableness value are peace-keepers, generally optimistic individuals, and trusting of others. They are friendly people and the low value on this trait is described as uncooperative individuals;
- Neuroticism: anxious, insecure, sensitive. Neurotic people are moody, tense, and easily tipped into experiencing negative emotions. People with a high value on this trait are very insecure, and the others that do not possess such an high value are more calm.

The work of U. Gupta and N. Chatterjee (2013) demonstrates the importance of the Big Five traits in the identification of human behaviour related traits trough psychological experiments such as deception, job performance, among other aspects.

2.2 Behaviour

Profiling an individual's personality can contribute to understand the potential needs in different contexts (Ortigosa, Quiroga and Carro 2011) and it is beneficial for many activities on a daily basis such as customer support, recommendation of services and products, and job applications (Poria et al. 2013). Personality is defined as the coherent patterning of affect, behaviour, cognition and desire over time and space, which are used to characterise unique individuals (Agarwal 2014).

Does personality provide us with unique explanations for human behaviour? The answer is 'Yes' (Higgins 2000). The belief is that personality-based variations in behaviour are largely interpretable in terms of the Big Five Factor traits (Paunonen 2003). Psychologists in general believe that personality affects various aspects of behaviour such as job performance, effectiveness, and dominance (Gupta and Chatterjee 2013). In fact, personality is an important trait that moderates people's behaviour and interactions with one another. For example, personality can be correlated with music taste, people with a high value on the Extroversion trait tend to like popular music while people that have a high value on the Openness to Experience trait tend to enjoy unpopular music. The Conscientiousness trait is a good predictor for overall job performance, job effort, and responsible work behaviour (Paunonen 2003). Extroverts are better at deception, Openness to Experience people tend to be more successful at work places, and an agreeable person is good at deception but he will seldom try to lie (Gupta and Chatterjee 2013).

Personality also has a strong correlation with emotion, a helpful analogy to understand this relation is to consider that personality is to emotion as climate is to weather. What we expect is personality, but what we can can observe in a particular moment in a particular context is emotion (Revelle and Scherer 2005). A person's behaviour is not simply a function of their personality traits, as an example, an aggressive person will behave aggressively in certain situations. The situational cues lead to activation of personality traits which then lead to a behavioural expression (Adali and J. Golbeck 2012).

2.2.1 Words As A Form Of Expression

Written language is one of the most primitive ways of communication, and similar to the way our actions and behaviour are directly related to our personality traits, even if people try to pretend being someone they are not, the influence of personality is so strong that they left some pieces of themselves on every sentence (Pitcher 2014).

The way individuals use words can reflect basic psychological processes, including clues to their thoughts, feelings, perceptions, and personality (Argamon et al. 2005). Different populations tend to write about different topics as well as to express themselves differently about the same topic. One simple communication task like e-mailing a friend about recent activities, is likely to be accomplished differently by two people. Some differences depend on their recent experiences, age, geographic location, past experiences, or on what they think interests the recipient, while others might depend on their character or personality (Oberlander and Gill 2006).

Some works (Marceau/Peyrouse 2009; Oberlander and Gill 2006) have already focused on email communications, by analysing the corpus of email messages in order to classify each user accordingly to the Big Five Model. They expected a more frequent usage of positive emotional language and social language, and more complex or extended expression from extrovert individuals (reflecting their tendency to dominate interactions). In the other hand, they were also anticipating a higher frequency usage of negative emotional language, self-oriented language, and more emphatic expressions, from neurotic individuals. Other work (Ding et al. 2015) also focused on email content, but the intention was to analyse the *phishing* susceptibility of a user by analysing their personality traits.

The work of Pitcher and Rod (2014) concluded that when analysing a text, either quantitatively or qualitatively, a lot of valuable and useful data is thrown away by ignoring words and phrases that are used figuratively, such as metaphors, exaggerations and pictorials. Some works approach this situation by content-based and style-based features (Marceau/Peyrouse 2009) or by analysing function words that may seem worthless but can actually tell a lot about someone (Pennebaker 2011).

Some characteristics that are likely to affect personality profiling include (Chin and Wright 2014):

- Word length of entries;
- Number of entries/author;
- Author identification;
- Spelling and grammar errors;
- Topic bias;
- Time-period bias;
- Author self-selection bias;
- Legal access and privacy restrictions;
- Unusual syntax usage, and abbreviations.

What about detecting the author of an anonymous text? This is an interesting topic approached by the work of S. Argamon et al. (2009) which show us that authorship profiling can help police identify characteristics of the perpetrator of a crime when there are too few (or too many) specific suspects to consider. In a similar way, large corporations may be interested in knowing what types of people like or dislike their products, based on analysis of blogs and online product reviews. Many different types of features have been considered as possible markers of textual style including lexical, syntactic, and vocabulary complexity-based features, but two basic features can be used for authorship profiling: content-based features and style-based features.

J. W. Pennebaker (2011) concluded that function words are important keys to someones psychological state and reveal much more than content words do. Pennebaker analysed the poems written by people who committed suicide versus poems by those who didn't, expecting to find more dark and negative content words in the suicides poetry. He didn't find that, but what he did discover was significant differences in the frequency of function words like 'I', and study after study he kept finding the same thing. When analysing military transcripts, he could tell people's relative ranks based on their speech patterns and it was the pronouns, articles, conjunctions, and other function words that made that possible. This is explained by the fact that in English vocabulary there are about 500 function words, and about 150 are really common. Content words (nouns, verbs, adjectives, and most adverbs) convey the guts of communication and they're how we express ideas, and help shape and shortcut language. Accordingly to the author, when the usage of function words is analysed it is possible to get a sense of peoples emotional state, personality, demographic aspects, and social class. Some demographic characteristics like age or gender can also be seen in the usage of function words. It is believed that men use the pronoun 'I' more frequently due being narcissists and self-congratulatory, however, across studies and cultures, Pennebaker work found that women use pronouns such as 'I', 'me', and 'mine' more often. Men use more articles ('a', 'an', and 'the') which means that men talk about objects and things. Women use more third-person pronouns ('he', 'she', and 'they') because women talk more about people and relationships, and they're better at managing them.

Another work (Marceau/Peyrouse 2009) indicate that neurotics tend to refer to themselves, use pronouns for subjects rather than as objects, use reflexive pronouns, and consider explicitly who benefits from some action. Non-neurotics, on the other hand, tend to be less concrete and to use less precise specification of objects or events (determiners and adjectives such as 'a' or 'little') and show more concern with how things are or should be done (via prepositions such as 'by' or 'with').

2.3 Conclusion

In this chapter it was possible to understand the ambiguity around the definition of personality. Personality takes a huge role in humans and it is a decisive factor that differentiates each one of us in such a way that our actions and patterns of behaviour are strongly connected to our personality traits, it is so strong that our own personality may take influence when answering the question 'what is personality?'.

It is possible to conclude that each one of the different theories on personality has their unique characteristics, since each focus on different visions. As a result, there is a well accepted and generally widespread model of personality classification that is based on the traits theory, the Big Five Factor

Model. This model of personality classifies each individual accordingly to five different traits of personality (which can have higher or lower values in each trait). The personality results from the combination of the values of each one of those traits.

The way we use our words can reflect basic psychological processes, including clues to our thoughts, feelings, perception, and personality. From demographic information, to overall interests and actions, it is possible to understand these characteristics by clues present on written text, either by content or style based features. This type of information, and the ability to profile the personality of an individual can be useful for security, criminal investigations, market research, target marketing, help the prevention of *phishing* attacks or even help creating adaptable user interfaces. It is important however to refer that this process is influenced by context, and differences in length and number of entries, syntax, abbreviations, spelling and grammar errors, and topics can take a huge influence.

Chapter 3

Ambient Intelligence

This chapter contains the vision for the interaction with environments instead of interacting with computers, known as Ambient Intelligence (AmI). This concept appeared at the end of the 20th century and intended to cause remarkable changes in the way people live. The initial part of this chapter contains a brief explanation of this concept, evidencing its origin, adaptation trough the times, and vision for the current times. The concept of ubiquitous computing, ubiquitous communication, and intelligent user friendly interfaces are also referenced on this chapter since they are considered three key technologies for AmI. The chapter proceeds by approaching the concept of Smart Environment (SmE), that is mutual to the concept of AmI, and contains a small overview for sensors, since they are a vital component and will be brought up in future chapters. Finally this chapter ends with an overview for daily applications of this concepts and technologies.

3.1 The Concept Of Ambient Intelligence

The immerse world of data that we live on today is full of possibilities and realisations of scenarios that were only possible on sci-fi movies or dreams. AmI and SmE are impulsed by ubiquitous computing and take advantage of the ease of collecting data from numerous devices present on our daily life.

At the end of the 20th century, M. Wiser et al. work (1999) described the existence of a new field of computer science created by ubiquitous computing, a field with a vision of a physical world filled with sensors, actuators, displays, and other computational elements, embedded on the objects of the daily life and connected through a continuous network. Ubiquitous computing is mainly driven by communication and following its definition (present, appearing, or found everywhere (Stevenson 2010)), means that we have access to computing devices anywhere in an integrated and coherent way (Ramos, Augusto and Shapiro 2008).

The concept of AmI was introduced by the European Commission's Information Society Technologies Advisory Group (ISTAG) (Ducatel et al. 2001) and this vision was based on the fact that at some point in time, humans will be surrounded by intelligent interfaces supported by computing and networking technology. This network of technology will be present on every aspect of our life, embedded in everyday objects, from furniture and clothes, to vehicles, roads, and materials. The focus of AmI concept have been adjusted in order to fit the chronological needs, as a matter of fact, in the 40's and 50's the attention was centred on the hardware, in the 60's that attention shifted to computers, 70's and 80's were focused

on networks, and from the 90's till the present day the attention is centred on the web (Ramos, Augusto and Shapiro 2008).

The report made by ISTAG (Ducatel et al. 2001) contained four scenarios with a vision of how AmI might be experienced in daily life and work around the year of 2010. The truth is that we have already passed that milestone, but these scenarios might sound familiar. These four scenarios emphasis the need for greater user-friendliness, more efficient services support, user-empowerment, and support for human interactions. Some key factors have resulted from these scenarios: social-political, business and industrial models, and technology.

We are involved with constant social and political decisions, and despise the fact that not everyone is willing to accept AmI on their lives, a series of necessary characteristics is needed in order to permit the eventual social acceptance. As result, an AmI should:

- Facilitate human contact;
- Be oriented towards community and cultural enhancement;
- Help to build knowledge and skills for work, better quality of work, citizenship and consumer choice;
- Inspire trust and confidence;
- Be consistent with long term sustainability and with lifelong learning;
- Be made easy to live with and controllable by ordinary people;

The potential business opportunities for AmI were identified as a job for the future generations of industrialists and entrepreneurs, and following that chronological line, that task is reserved for the industrialists and entrepreneurs of today. Some aspects were however referenced:

- Initial premium value niche markets in industrial, commercial or public applications where enhanced interfaces are needed to support human performance in fast moving or delicate situations;
- Start-up and spin-off opportunities from identifying potential service requirements and putting the services together that meet these new needs;
- High access-low entry cost based on a loss leadership model in order to create economies of scale;
- Audience or customer's attention economy as a basis for 'free' end-user services paid for by advertising or complementary services or goods;
- Self-provision (based upon the network economies of very large user communities providing information as a gift or at near zero cost).

In terms of technology needs to make AmI a global reality, the report identified five requirements:

1. Very unobtrusive hardware;
2. A seamless mobile/fixed communications infrastructure;
3. Dynamic and massively distributed device networks;
4. Natural feeling human interfaces;

5. Dependability and security.

Context awareness is one of the most desired concepts to include in AmI, the identification of the context is important for deciding to act in an intelligent way (Ramos, Augusto and Shapiro 2008). We face the need to successfully identify human needs, personality and behaviour in order to prepare appropriate interactions when facing different types of events, and this new paradigm aims to discover information about the environment state in order to fulfil those necessities (Cottone, Re et al. 2013). AmI is aware of the specific characteristics of human presence and is able to take care of needs, responding intelligently to spoken or gestured indications, and can even engage in intelligent dialogue. This process should also be unobtrusive (often invisible) and the interaction should be relaxing and enjoyable for the human, with the most natural feeling possible, and not involve a steep learning curve (Ducatel et al. 2001). In order to achieve this vision intelligence must be provided to our environment (either in the context of intelligent homes, intelligent vehicles, or even intelligent cities), and for this reason this concept is not possible without Artificial Intelligence (AI). As a result, AI researchers must be aware of the need to integrate their techniques with other scientific communities techniques (automation, communication, machine learning, computational intelligence, natural language, knowledge representation, computer vision, intelligent robotics) (Ramos, Augusto and Shapiro 2008).

In essence, AmI results from the convergence of three key technologies: ubiquitous computing, ubiquitous communication, and intelligent user friendly interfaces.

3.1.1 Ubiquitous Computing

The general definition of ubiquitous computing technology is the continuous and discrete presence of computational systems that liberate people from a large extent of tedious routine tasks. All models for ubiquitous computing share a vision of small, inexpensive, and robust networked processing devices, that are distributed on everyday aspect of our daily life. This means that any computing device can build incrementally dynamic models of various environments, with the capability for recognise past environments they have operated in, or proactively build up new services and environments (Raisinghani et al. 2006).

3.1.2 Ubiquitous Communication

Ubiquitous computing is not a strange term for all of us, in the current time, numerous objects are equipped with computers. From the smartphone we carry on our pocket, to the car we drive, or to our household appliances, all of those scenarios are equipped with computers but, in most cases, the computers do not operate at their full potential since they are unable to communicate with each other.

Ubiquitous communication refers to the capability of accessing networks and services from anywhere, and the introduction and expansion of wireless network technology enables flexible communication between interlinked devices. However, the mere existence of wireless technologies does not suffice to promote ubiquitous communication and computing. It is vital to assure network integration, communication, and security, and in order to combine computers and networks efficiently and effectively it is

crucial to have a communication without the necessity for data conversion or translation. The combination of the above characteristics is referred as network interoperability and it is a imperative resource for the success of AmI (Raisinghani et al. 2006).

3.1.3 Intelligent User Friendly Interfaces

Intelligent user friendly interfaces, or user adaptive interfaces, goes beyond the interactions by the traditional keyboard and mouse, to improve human interaction with technology by making it more intuitive, efficient, secure, and shifting around their users preferences. This interfaces allow the computer to know more about an individual, about the context, the environment, and related objects that can be interacted with (Raisinghani et al. 2006).

This intelligent social user interfaces can be grouped into five categories:

1. Visual recognition (face, 3D gesture, and location) and output;
2. Sound recognition (speech, melody) and output;
3. Scent recognition and output;
4. Tactile recognition and output;
5. Other sensor technologies.

Some of these categories may sound familiar since they are already present on our smartphones that we carry everyday, the virtual assistants that help us with some tasks (such as smartphone assistants or smart speakers), and other numerous devices that, for example, make use of visual and sound recognition.

3.2 Smart Environments

AmI enhances the global behaviour of a system by providing high level functionality, which provides an added value to the typical services expected in a specific environment, and SmE are linked to this same concept and vision. These two mutually complementary areas are growing together and have the same vision to benefit society. A SmE is an ecosystem of interacting objects that has been enriched with technology (sensors, processors, actuators, information terminals, and other devices interconnected through a network) that have the capability to self-organise, provide services and manipulate complex data. This physical space is smart in nature, and that smartness results from a interaction of different devices and computing systems, and aims to enhance the services that can provide to humans.

Similarly to AmI, there were identified three computational areas that must converge in order to develop a truly SmE: ubiquitous computing, intelligent systems, context awareness. Ubiquitous computing is responsible for providing a seamless interface between the environment and its users, and the integration of the system into the everyday objects must be simple, natural, and a non intrusive experience for the users. Intelligent systems are responsible for inferring the context of the environment and understand patterns based on the users behaviour (techniques such as data mining, statistical analysis, machine learning, or optimisation methods, take a huge role on this scenario). Context awareness is the adaptation

of the environment to their users habits. This requirement is vital in order to have a system that shifts around their user needs by perceiving the context (using sensors) and capable to change around it (using actuators) (Antunes, Gomes and Aguiar 2013).

There is a wide range of assistance that an SmE system can aim to provide, and despise the common association with "Smart Homes", SmE are equally applicable to hospitals, cars, offices, streets, and a higher range of environments. In the context of smart homes, typical examples are actions aimed at preserving or increasing safety, encourage better life styles by comparing trends on the activities developed over a long period, or even facilitate and aid some tasks. By focusing on the rational use of energy in order to save money and be environmental friendly, to providing personalised television or music services with the scope of entertainment in mind the concept of smart homes is appealing and it is no surprise that this is the scenario most associated with SmE. Equally, cars can be transformed into SmE to assist drivers in difficult conditions (or even engaging auto-pilot mode), and other social scenarios like classrooms can benefit from SmE by being equipped to enhance the teaching-learning experience, as well as offices can be supplemented with technology to support effective work-group collaboration.

Independent of the context, usually it is performed an analysis in real-time over the events that are recorded within the SmE which allows a timely interaction with the inhabitants of the environment to provide a service.

3.2.1 Sensors

In order to fulfil the necessities for context awareness for perceiving the environment, sensors are a principle fundamental for SmE.

The concept of AmI or SmE is often associated with intelligent sensors that are embedded in our environment. This physical devices are usually conceived for detecting or measuring motion, light, temperature, humidity, and other conditions that are descriptive of the environment (Cook and Song 2009).

The role of sensor networks in an AmI environment is to furnish the higher levels of the system with answers to the following questions (Pauwels, Salah and Tavenard 2007):

- Who? by tracking and identifying persons and/or pets;
- Where and When? by providing a time frame for location and object associations in order to determine context:
- What? by recognising activities, interactions, relations, as well as linguistic and non-linguistic messages, signals, and signs;
- Why? by associating actions with semantics, plans, task identification, and recognising behavioural patterns;
- How? by tracing the information flow trough multiple modalities, recognising expressions, movements, and gestures.

The problem of activity recognition has been addressed by the usage of intrusive sensors (like wearable sensors) positioned directly or indirectly on the body, embedded into clothes, eyeglasses, belts, shoes,

wristwatches, and others (Cook and Song 2009). But the release of technologies such as the Microsoft's Kinect sensor, allowed researchers to perform activity recognition on an unobtrusive perspective (Cottone, Re et al. 2013).

3.3 Daily Applications

Major opportunities to create an integrated AmI landscape can be built upon European technological strengths in areas such as mobile communications, portable devices, systems integration, embedded computing and intelligent systems design (Ducatel et al. 2001). The implications for issues such as energy, environment, social sustainability, privacy, social robustness and fault tolerance may in the longer run determine the success or failure of AmI. Despise the common association with smart homes, AmI and SmE have been used to produce tasks such as optimisation of energy consumption (Cristani, Karafili and Tomazzoli 2015; Stavropoulos et al. 2015), recognition of human activity and preferences (Cottone, Maida and Morana 2014; Cottone, Re et al. 2013), aid the elderly or persons with health problems (Ramos, Augusto and Shapiro 2008), or even increase the lifestyle of blind people (Mekhalfi et al. 2016).

A recent application that aims to facilitate human interaction with the environment, is the case of the Amazon's Echo. It can do a lot more than answer a question, including keep track of a shopping list and place orders, book an Uber ride, control a thermostat and other household appliances, tell you transit schedules, start a seven-minute workout routine, read recipes and do math. Among this features, it can even call a plumber and share medical advice (Grossman 2016).

3.4 Conclusion

This chapter introduced the concept of AmI, the correlation with SmE, their key components, and daily applications that are already implemented. This concept is establishing as an area that aims to help society trough technology. From the general idealisation of smart homes, to applications in health care and recognition of human behaviour, it is possible to verify the presence of this concept in our daily lives. Despise the many challenges and improvements needed, namely in the ubiquitous communication and intelligent user friendly interfaces, this concept is growing and evolving at a fast pace.

There is the need for context awareness, since SmE aims to successfully identify human needs and preferences, and needs to shift around them. In order to achieve the desired result, SmE make use of physical sensors that need to answer the questions 'Who?', 'Where and When?', 'What?', 'Why?', 'How?'.

Technology should enhance the quality in life and facilitate daily activities, and it is already possible to experience the implementation of AmI and SmE on the daily bases. From the car that learns your driving patterns in order to save fuel and increase efficiency, to the virtual personal assistants that aid us with our tasks and we can communicate with, this concept is already imposing their position in modern society, becoming less and less intrusive, and starting to become a normative aspect in our life.

Chapter 4

Online Social Networks

The present chapter is focused on the Internet and its relation with the tasks present on our daily lives as well as a social aspect that is embedded into the reasons that lead people to use different services and platforms. This chapter contains an overall vision about the growth of Internet and the consequences of it. Online platforms like Online Social Networks (OSNs) have increased in a fast rhythm, and this growth is motivated by the offers and opportunities associated with each different OSN. The motivations directly related related to the growing numbers of OSN usage and content, as well as the motivations that driven people to use those platforms, are divided into two main categories (the need to belong and the need for self presentation), that are also described in this chapter.

4.1 Internet Trough Numbers

Since last decades that the growing of the Internet, and its usage, reached an impressive rate. Is estimated that at the year of 1995 the percentage of people in the world with Internet connection was less than 1%. This value is way different when talking about our current year (2017), were is estimated that the percentage of people with Internet connection is around 46%, and is increasing day after day (InternetLiveStats 2017).

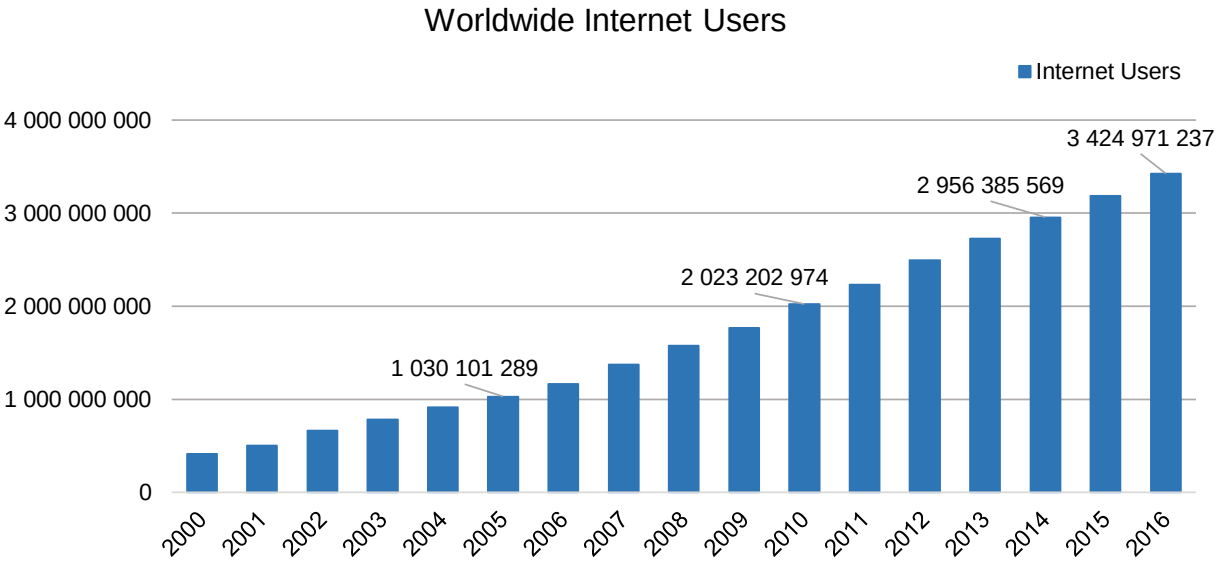


FIGURE 4.1: The number of global Internet users per year since 2000

By looking at the data present on figure 4.1 it is possible to understand the increase of the Internet usage across the years. The milestone of 1000 million users was reached in 2005 and five years later was reached the milestone of 2000 million users. The current milestone of 3000 million users was reached at the end of the year 2014, and only after past a few years, we are already at half way to the 4000 million users mark (estimated value for 1st July 2016). The decrease on the interval value between milestones may be an indicator that we are going to achieve the next milestone in less time that we needed for previous milestones.

The growth of Internet users across the globe also increased the growth of numerous platforms and services such as online social networks (OSNs). An OSN can be defined as a network of social interactions and personal relationships, and a platform which enables users to communicate with each other by posting information, comments, messages, images, among other communication channels (Stevenson 2010).

4.1.1 The Growth Of Online Social Networks

A large number of people shifted themselves into the virtual world making their online virtual profile a mirror of their true identity. It is unrealistic to say that this is a transient trend when the numbers speak for themselves, as observable in figure 4.2. The numbers of January 2017 gave the market lead to the popular OSN Facebook with an astounding number of 1871 million active users, being the first social network to surpass the mark of 1000 million registered accounts. This number is followed by the OSNs Facebook Messenger and WhatsApp with 1000 million active users each, QQ (Chinese social media network) with 877 million, Instagram achieving the 600 million mark, and a little bellow is possible to find OSNs like Tumblr and Twitter with 550 and 317 million active users respectively (Statista 2017a).

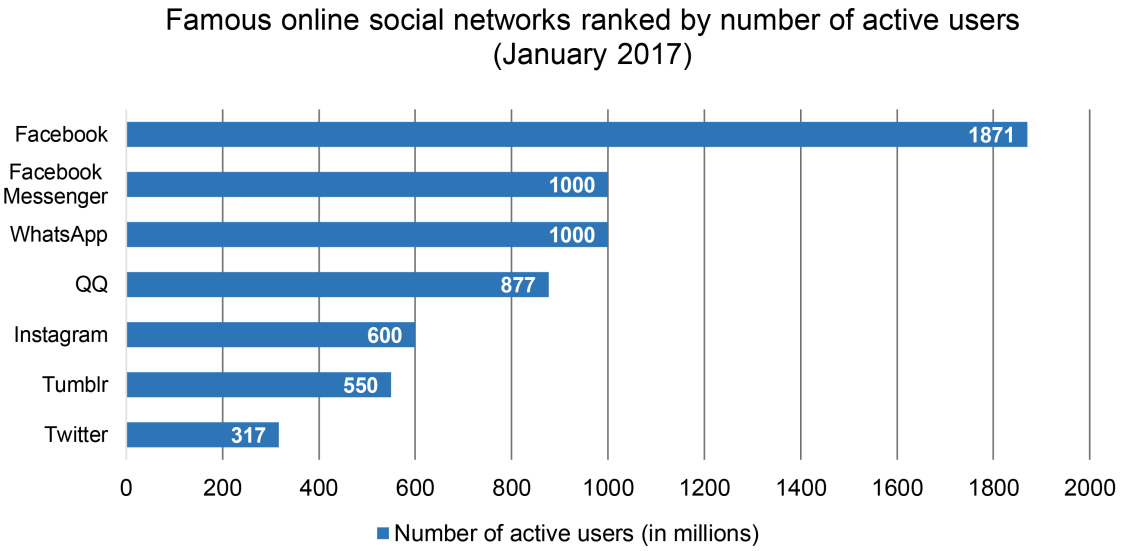


FIGURE 4.2: Famous social network sites worldwide as of january 2017, ranked by number of active users (in millions)

The growth of Internet usage is also reflected in a growth on the amount of social media users. Figure 4.3 contains a representation of social media growth from year 2010 until the current year (2017), as well as

a projection for the following years. The data sugests that in 2020, the amount of social media users will be close to the 3,000 million milestone (Statista 2017b).

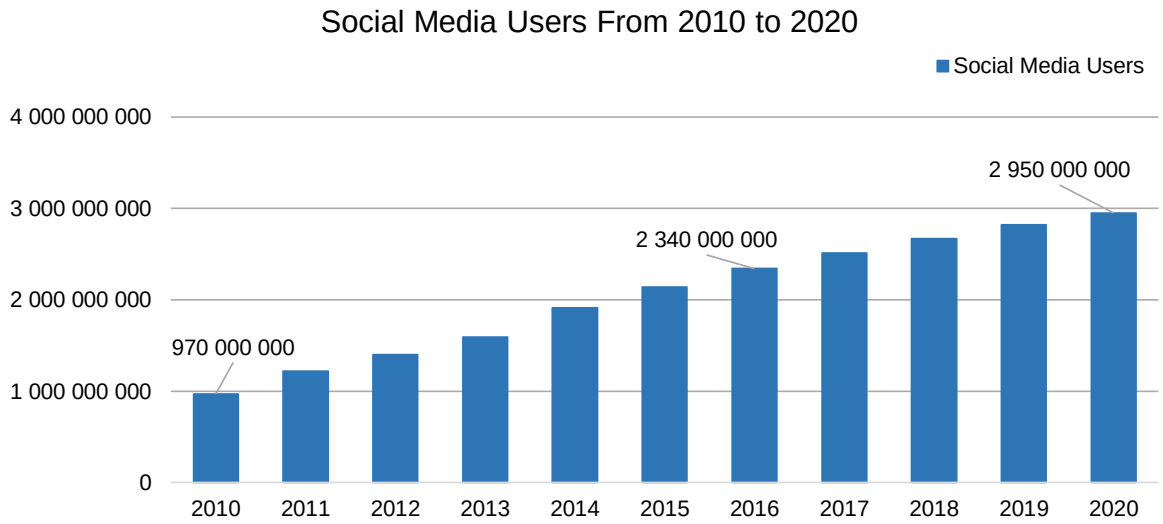


FIGURE 4.3: Growth of social media users worldwide from 2010 to 2020 (in millions)

OSNs have become so popular that it is estimated that about 68.3% of Internet users access a social network regularly on a daily basis and have an average of five social media accounts (Smith 2016). Age is strongly correlated with social media usage, young adults (ages 16-24) are the most likely to use social media by a considerable margin, but the usage among older users (65 years old and beyond) have more than tripled since the year 2010 (Shannon Greenwood and Duggan 2016). The OSNs popularity is spread across different characteristics, as an example, OSN Pinterest has a particular appreciation among female users, LinkedIn is especially popular for newcomers, and Tumblr or Instagram have particular interest from young users (although this last two OSN have a substantial overlap between their users) (eMarketer 2016).

As seen in figure 4.4 (Chaffey 2016), the OSN usage is similar across different age groups, only with the exception of Tumblr or Instagram that are really popular among younger age groups. This similarity shows that OSNs are now at a stage of maturity where they give opportunities to reach all age and gender groups. When talking about messaging tools, it is possible to observe the same similarity across age groups, with the exception of Snapchat, Kik, and Wechat, which are clearly more popular with younger age groups.

Using for example the OSN Facebook (selected as an example due being, until today, the OSN with higher values of preference among users), is estimated that at each 60 seconds new 510 comments are made, 293,000 profile status are updated, and are uploaded 136,000 photographs (Zephoria 2016). On average, the time spent per Facebook visit is 20 minutes, the 'Like' and 'Share' buttons are viewed across 10 million websites daily, and a total of 300 million photographs are uploaded each day. All of this growth also generated a growth in data that is created everyday. Looking exclusively at the unstructured data present in each comment that is made, it is possible to say that at each hour we have 30,600 new sources of information in the format of natural language, 734,400 each day, which totals for 22 million each month and more than 2,500 million each year. This enormous amount of data can be used to produce knowledge about an individual, and for that reason OSN are appealing to organisations.

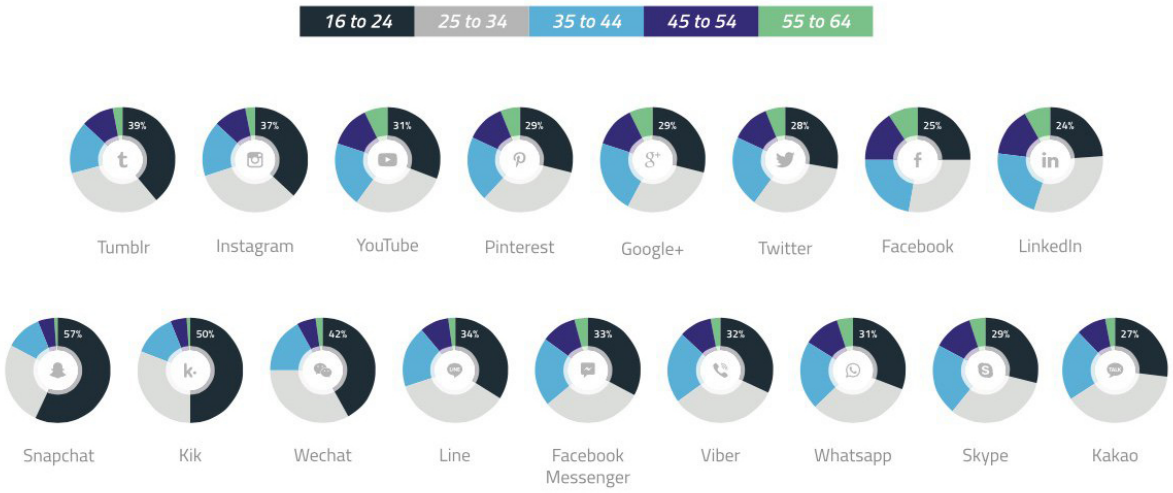


FIGURE 4.4: Active user age groups of the top social platforms and messaging tools (Chaffey 2016)

In fact, due to the impact on the daily life of people, even companies and brands are using OSNs. It is normal to see business cards, or even products, emphasis the organisation social media pages, sometimes even in function of their own websites. Organisations have shifted some of their focus into OSNs as a channel of communication, marketing, approximation to their customers, and identify and reach new potential customers. OSNs have become a valuable asset for organisations since they are a great starting point for any business, they can be used to share media content (photographs, videos, and others), share important company updates, and require less maintenance compared to a company website.

4.2 Why Do People Use Online Social Networks?

Due to a constant presence in the lives of their users, OSNs have a decidedly strong social impact leading to a blur between offline and virtual life as well as the concept of digital identity and the motivations for their usage differ from person to person, while some focus on broadcasting information about themselves others are more interested in passively consuming information produced by others (Wald, Khoshgoftaar and Sumner 2012).

OSNs not only permit users to socialise with others, but also offers the possibility to construct and manage their identities (Lee, Ahn and Y. J. Kim 2014), by creating their visible profiles where is required, at a minimum, a name, gender, and a date of birth. Among these required basic fields, users can add basic facts about themselves such as home town, contact information, personal interests, job information, and even a profile photograph.

As seen in figure 4.5, the most common reasons for using an OSN are all related. Those reasons are related to what most OSNs offer to their users, a way to stay in touch with friends, to stay up-to-date with news and current events, and even to fill up spare time. Among those reasons, there is also motivation related to the expression of opinions, content consuming and sharing, or even share information about daily life events (McGrath 2015).

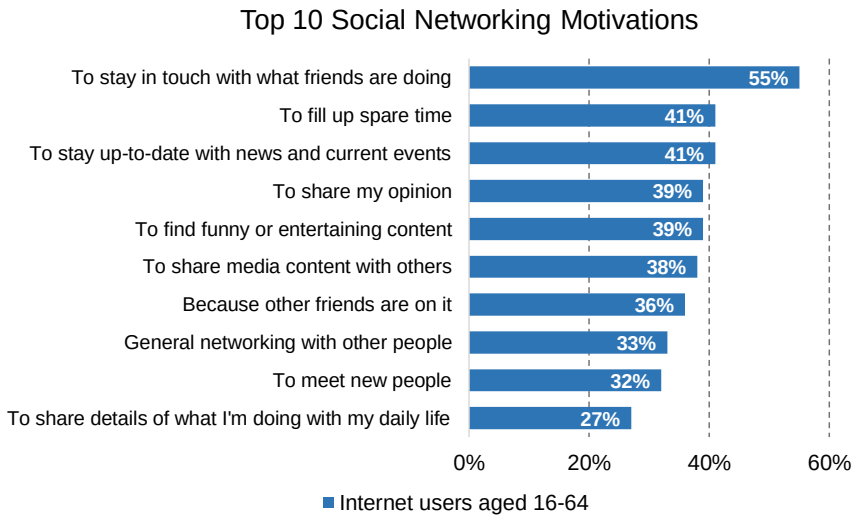


FIGURE 4.5: Top 10 reasons to use social media by Internet users aged 16-64

Through the usage of ONSs individuals often express preferences for brands, products, services, persons, or even political inclination, in a free unsolicited way (Dam and Velden 2015). OSNs connect people who share interests and activities across geographic borders and have become a virtual mirror where users reveal a lot about themselves both in the way they share information and how they share it.

OSNs request that users construct truthful representations of themselves with varying degrees of accuracy (Amichai-Hamburger and Vinitzky 2010). Even demographic aspects can influence the type and frequency of usage, as an example, in their work K. Moore and J. C. McElroy (2012) found a significant positive relationship between gender and a number of variables of interest where was possible to find that women spend more time on the OSN Facebook, had a greater number of friends, posted more photographs, and did more postings about themselves, when compared to men.

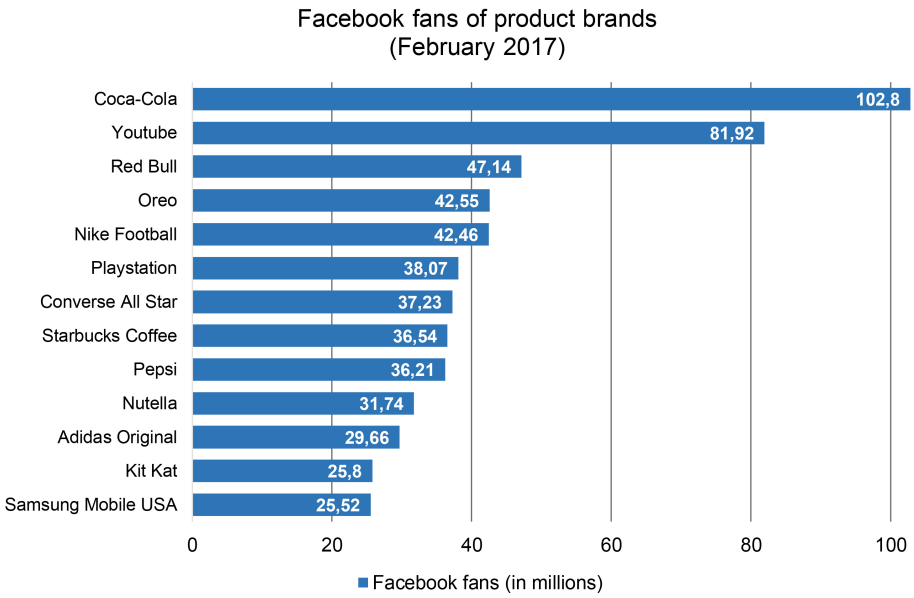


FIGURE 4.6: Product brands with the most Facebook fans as of february 2017 (in millions)

Following the tendency of their customers, even organisations became more interested in OSNs. When

users aged from 18 to 34 are most likely to follow a brand or business social page, and 71% of those users have the intention to recommend those pages to others, the strategy of the organisations started to include OSNs. If done correctly, the presence on social media can increase business reach, obtain more exposure for a brand, increase the service and product exposure, and ultimately increase leads, customers, sales, and revenue (Santoro 2016). In fact, the data of February 2017 represented in figure 4.6, evidence the incredible success that some organisations are having already on the popular social network Facebook, with Coca-Cola surpassing the mark of 100 million fans. Interesting fact, even other OSNs (like the case of Youtube) are using others social networks in their marketeer strategy. This is not a surprise, since usually, an organisation has at least two different OSN profiles because they want to reach the most people possible.

Accordingly to Nadakarni and Hofmann model (Nadkarni and Hofmann 2012) the motivation for the usage of a OSN is primarily motivated by two basic social needs, the need to belong and the need for self-presentation.

4.2.1 The Need To Belong

The need to belong is associated with the necessity for affiliation with others and the gain of social acceptance, since humans are highly dependent on the social support of others. Some type of obstruction from the social group has a negative impact in humans on a variety of health-related variables, including, self-esteem and sense of belonging, emotional well-being, sense of life meaning, purpose, self-efficacy, and self-worth (Nadkarni and Hofmann 2012). In fact, the need to belong is highly related to the following characteristics (Selterman 2012):

- Creating social bonds: people quickly form relationships with others without being paid or forced to do so, and this happens even under adverse circumstances. For example, infants and children will form attachments to others even though they have little or no knowledge of their social world, and are incapable of calculating benefits or costs to those relationships;
- Not breaking bonds: people are eager to have close relationships and are reluctant to break them once formed, even when the relationship is marked by distress, conflict, or even abuse. People often avoid permanent separation (breakups, divorce, death), even when the costs of staying in the relationship are greater than leaving;
- Cognition: when feeling close to others, people thoughts change such that a cognitive "merging" effect occurs (people begin to include aspects of their relationship partner in their own self-concept). The boundaries between individual partners break down, and people think of their own fate as being crossed with the fate of others;
- Emotional highs and lows: relationships carry immense emotional weight, and people feel a great deal of positive emotion (joy, bliss, love), especially during the early stages of relationships. People also feel lots of negative emotions and distress (anxiety, anger, jealousy) when there is a problem associated with the relationship;

- Consequences of deprivation: close relationships boost people's immune system. When people lack meaningful close relationships with others, people often suffer. Specifically, married individuals are healthier, less stressed out, and are expected to live longer than single individuals;
- Partial deprivation: even within highly satisfying relationships, being separated from a loved one (or having restricted interactions) produces distress and sadness;
- Satiation and substitution: people strongly prefer to have (and are only capable of having) a few very close friendships and a larger number of casual friendships. In this case, quality is more important than quantity. Relationships take time, effort, energy, and resources, so it makes sense that any individual person would experience a "satiation point" after their needs are fulfilled. In addition, when a bond is broken, people will readily pursue another in its place.

Seidman (2013) says that the need to belong is a fundamental drive to form and maintain relationships and a major motivator factor for OSN use. OSNs like Facebook allows users to fulfil belonging needs through communicating with and learning about others. Facebook can be an effective method for coping with feelings of social disconnection, as it enables peer acceptance, relationship development, and can even boost self-esteem.

4.2.2 The Need For Self-Presentation

The need for self-presentation is correlated with the continuous process of impression management. OSNs opens the possibility for its users to display their idealised, rather than accurate, selves through their profiles (Nadkarni and Hofmann 2012).

This need is also related to impression management, which means that people experience a process of portraying themselves in a manner that creates the desire impression. In fact, individuals attempt to control, or guide, others impressions by changing or fixing their appearance or manner. Every individual is engaged in certain practises to avoid being embarrassed or embarrassing others. This impression management is neither good nor bad, it is an integral part of our social interaction and everybody gets involved in it everyday.

Activities that accomplish self-presentational goals include posting photographs, profile information, and display relations. According to Seidman (2013) popularity seeking users tend to disclose information, engage in strategic self-presentation, and enhance their profiles (that generally represent an accurate self-presentation).

4.3 Conclusion

The growth of the Internet and its platforms, more in specific the OSNs, have attracted large amounts of individuals. Either motivated by the need to belong (the association with others and the seek for social acceptance) or the need for self-presentation (the continuous process of impression management), different types of OSN platforms attract different type of individuals, which can lead to a different demographic presence across them.

The possibility to stay in touch with friends, to stay up-to-date with news and current events, construct a virtual identity, or simply fill up spare time, makes OSNs appealing for their users that have included them into their routines on a daily basis. This heavily presence of users that spend a considerable amount of time using OSNs on a daily basis results in a presence of personal information, in addiction to demographic characteristics, that is expressed by their preferences for brands, products, services, persons, or even political inclination.

On other perspective, organisations have followed the tendencies of their customers and adopted OSNs as a channel of communication, marketing platform, obtain insights about their current customers, or even empower the discover of new customers or new market opportunities.

Chapter 5

The Dimensions Of Complex Networks

The constant presence of different types of networks in the real world have motivated the study of their representation, known as complex networks. From social interactions, to the different types of transportation methods that compose a transportation network of a city, to the power line networks or even all of the utility networks that are present on our homes, almost every single aspect of our world can be represented by a network.

This chapter emphasis the study of complex networks and the motivations associated to it. It starts by referencing the structure and the existence of complex networks, and then evidences the need for multidimensional approaches that resemble the real world networks. The chapter ends with a reference to a multilayer approach that has been proposed to obtain even more insights and perspectives when representing real world scenarios into complex networks.

5.1 A World Of Networks

The dramatic increase of interest in network models has been determined by the wide applicability of these models to several disciplines, including social network analysis, biology, physics and economics. The interdisciplinary character of this problem has influenced recent developments in the field of computer science, where simple graph models have been enriched with additional information to conform to the representation needs required in other disciplines (Magnani, Monreale et al. 2013).

In fact, our actions and interactions present on a daily basis can be represented in a network. It is possible to observe a large number of interactions and connections among information sources, events, people, or items, that are known as complex networks. A complex network is represented by a graph with a set of nodes connected by edges, that together form a network. In this representation, nodes represent entities and the edges represent the relation between them (J. Kim and Wilhelm 2008). Multidisciplinary and extensive research has been devoted to the extraction of non trivial knowledge from this networks. Complex network analysis is associated with, for example, prediction of future relations among actors of a network, detecting and studying the diffusion of information among them, or even mining user behavioural patterns (Berlingerio et al. 2013).

Enumerating all the possible networks detectable within our world, or their properties, would be difficult due to their number and heterogeneity, however, the work of Newman (2003) classifies this networks into the following categories:

- Social: networks related to the social aspect and the relations between individuals. This networks are found in OSNs, as well in the definition of interactions among individuals in the physical world;
- Informational: networks that describe the relationship between information such as citation networks, or online encyclopedias;
- Technological: examples of this type of networks are found in power grid networks, transportation routes (such as airlines, trains, buses, and others), the Internet, or even some structural networks like cities or the pipes present in a habitation;
- Biological: this networks are used in the representation of biological characteristics, like protein interaction networks, or even the representation of the human brain.

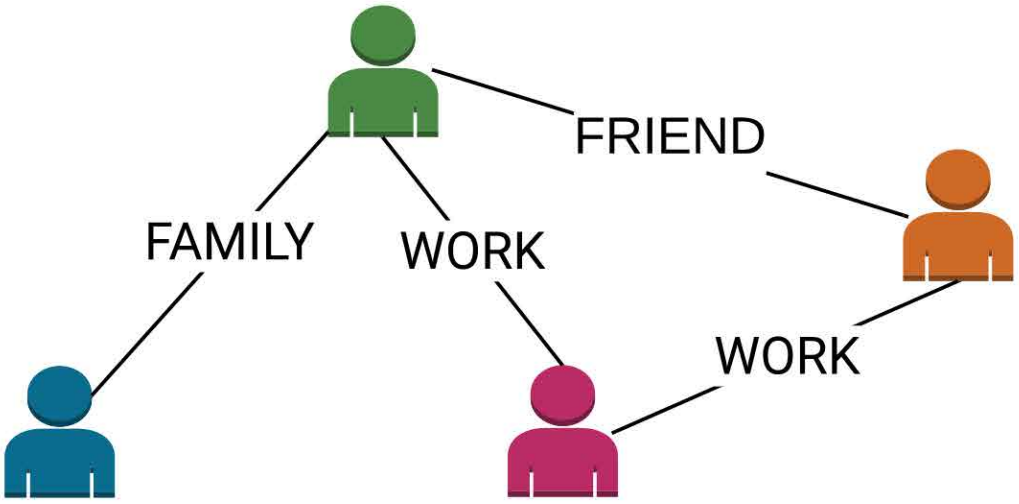


FIGURE 5.1: Example of a monodimensional complex network

As a visualisation of this concept, the figure 5.1 contains a representation of a complex network categorised, accordingly to Newman (2003), as a social complex network. In this network are represented, in a very small scale, actors (in this case persons) and the social relations between them. In the context of social networks, accordingly to Magnani et al. (Magnani, Monreale et al. 2013), is important to consider the differentiation between several types of relations, and in example it is possible to observe that the actors (the nodes) are connected by a social relation (the edges). If it is family bounds, sharing a working place, being friends, or any other type of relation, they all can be represented in a complex network.

One important aspect related to this type of networks is the fact that they are classified as monodimensional complex networks, meaning that can only exist one relation between two nodes.

5.2 Multidimensional Networks

Despite the current possibility of representing different scenarios into monodimensional complex networks, in the real world, networks are often multidimensional, meaning that there might be multiple connections (relations) between a pair of nodes. The existence of those different relations between to

nodes is referred as dimensions, and different dimensions may reflect different types of relationships, or even different degrees and values of the same relationship (Berlingerio et al. 2013). Therefore, multidimensional analysis is needed to create a distinction among different types of relations.

In the real world is possible to observe those multidimensional characteristics in numerous networks such as:

- Transportation Networks: looking at a transportation network present in a city, an entire country, or even in the world, it is possible to observe a multidimensional network where cities are represented as nodes, and each transportation mean is a dimension. By addressing these components into this perspective it is possible to observe that each city is connected to other cities by a transportation method such as aeroplanes, trains, ferries, or any other transportation mean. But, as easily imaginable, it is possible to connect different cities by more than one method of transportation (cities A and B can be connected by both train and aeroplane). One interesting feature related to this type of networks is the multidimensionality associated with each travel plan, meaning that even if the travel is meant to be using aeroplanes, at some point in time other dimensions like buses or trains are used (in order to reach the airport). The context of this networks can be shortened to a point where the entire network is related to a method of transportation and the dimensions are associated with transportation companies (for example, when looking at a network of cities connected by aeroplanes, the dimensions can be a representation of the different airlines that provide those inter-city connections);

- Social Network: when addressing the social networks present on the real world, it is easy to visualise different connections between persons, in fact, a person can be related to other by a friendship relation, and share the same gym or workspace at the same time. Those multiple relations (dimensions) are part of the backbone associated with the current popular OSNs, where they provide different dimensions between users and even with an associated degree value (in the case of the OSN 'LinkedIn' where each relation is defined by a first, second, or third degree). Despite not being new, the concept of multidimensional network is fairly recent in the scope of OSNs (Kivela et al. 2014). Socievole et al. work (2015) makes reference to the effort that has been made into the definition of multiple social metrics that consider all the existing different social dimensions. Other works have attempted to define models (Bródka, Przemysław Kazienko et al. 2012; Magnani and Rossi 2011; Socievole et al. 2015), and even looking at relations derived from structure and written text content present on OSNs (Forestier, Velcin and Zighed 2011). It is important to notice that when talking about multidimensional approaches in the context of OSNs, those works are heavily focused on the social interactions between users (exploring the need to belong). There is an opportunity in the application of this concept, not only for the social relationships, but as well with the intention to explore the different degrees of interests of those users;

- Utility Network: networks such as power line networks, water pipes, telephone and television cables, all belong to a group of dimensions that are present, for example, in each house or office network. In the context of a house or office, the node level of the multidimensional network is highly redundant since almost every node is served by every utility, but the network structure (like the distribution of the relations) might differ. This is a great example to demonstrate how many different dimensions can exist in the same context and even complement other dimensions.

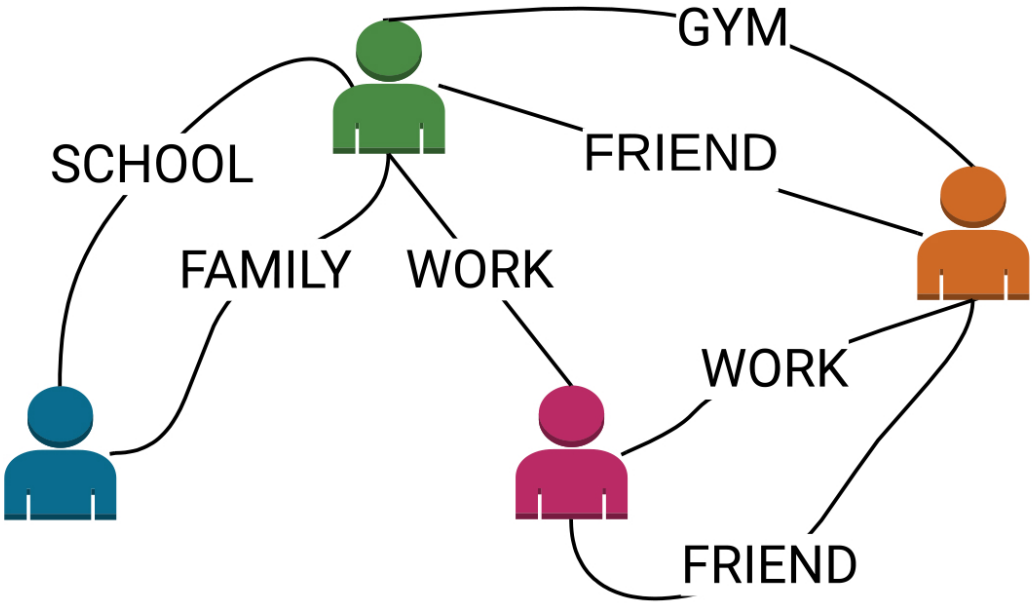


FIGURE 5.2: Example of a multidimensional complex network

There is little doubt that the multidimensional approaches are a significant oversimplification of the rich complexity that exists in most of the networks present in the real world. As a general analytic system, network analysis can be applied to both an amazingly diverse set of objects and a similarly diverse set of relations (Contractor, Monge and Leonardi 2011). Despite the correct representation of a real world social network illustrated by the figure 5.1, when looking at those real life case scenarios, the presence of different relations between two persons is easily perceptible. In fact, real world social networks (among other types of networks described before) often contain a multiplicity of relations, that are best represented in a multidimensional network. For example, a simple transportation network like the ones found in every city is not populated by a single method of transportation, in fact, every city contains several transportation methods that co-exist in the same transportation network. When looking at real life social networks the same principle is applied, and as described by figure 5.2, despise having a main connection between them, people are often connected by more than a single relation. Either by family relationship, sharing workplaces, attending to the same places, between many others relations, it is natural that when representing real life scenarios into complex networks, to consider the multiplicity of relations present on the real world and include them in a virtual network.

5.3 Multilayer Networks

With multidimensionality being a characteristic of our world it is only natural that multidimensionality is reflected and represented in complex networks, despite the lack of consensus towards the terminology (where terms such as multi-relational network, multidimensional network, and multiplex network are considered synonyms (Bródka and Przemyslaw Kazienko 2014)).

Even if multidimensional networks can represent real world networks with diverse degrees of accuracy, there is more to be added to this topic. Kivelä study on multidimensional networks (2014) emphasis the

increase on the study of networks composed by multiple layers. This study addresses multidimensional networks on a multi-layer perspective where each layer is composed single dimensions or a set of dimensions. Looking, for example, at a transportation network that contain various dimensions (depending on the number of relations between the different transportation methods), this multi-layer representation will consider each transportation method into each layer, or even represent each transportation type in each layer (for example, one layer can be dedicated to air travel and can contain helicopters, airline companies, private aeroplanes, among others).

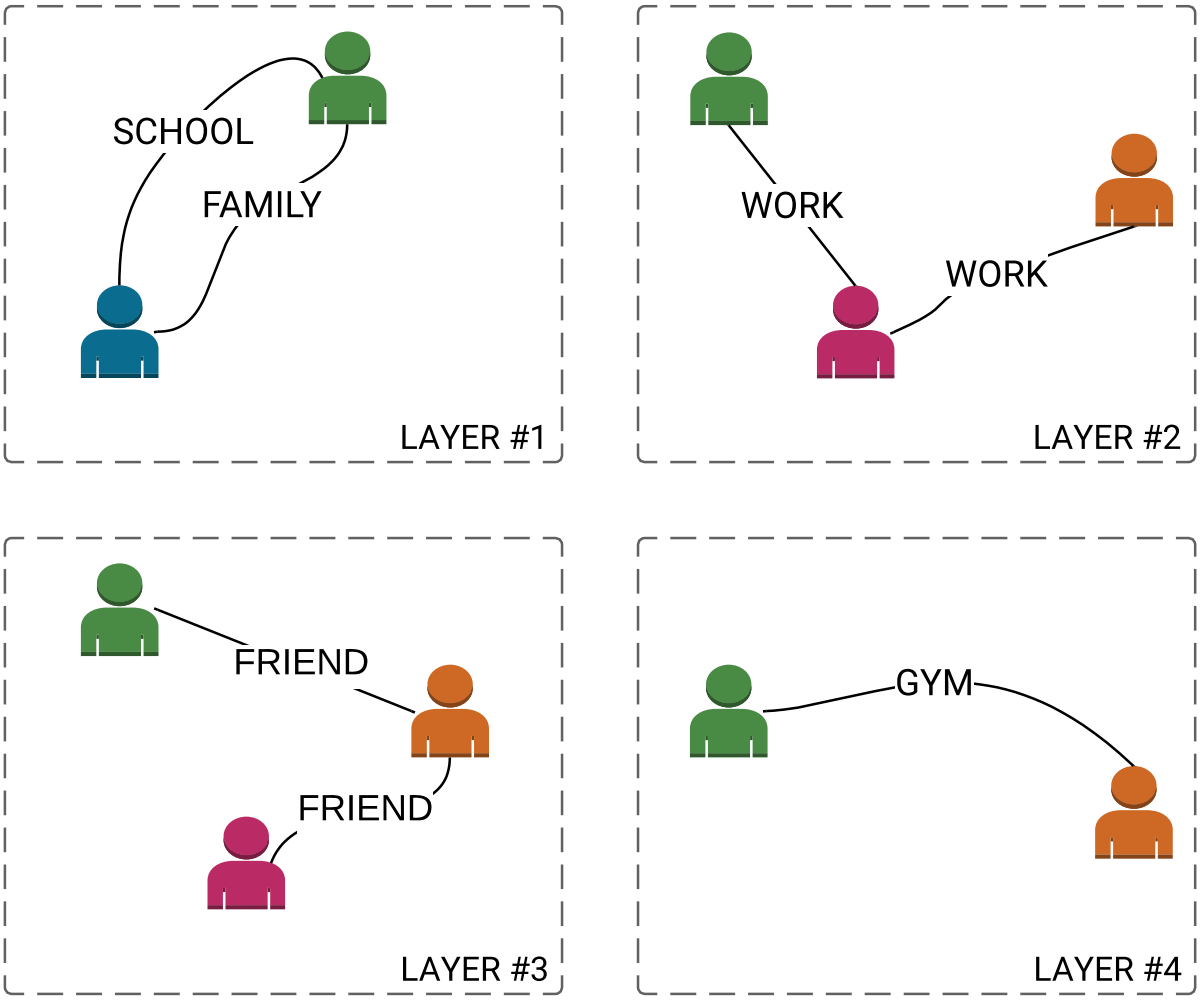


FIGURE 5.3: Multidimensional network represented by layers

Recapping the previous examples represented by the monodimensional complex network on figure 5.1 or by the multidimensional "evolution" represented by figure 5.2, in the context of real world social networks, it is possible to represent dimensions, or sets of dimensions, into layers. The multi-layer representation present on figure 5.3 addresses that characteristic by representing relations between persons in different layers. Each layer, in this example, is composed by relationship dimensions between persons and can be defined as some type of "filter" that helps to visualise and combine different dimensions. With this multi-layer representation it is possible to combine dimensions accordingly to the interests of the subject or the context that is given to that network (for example in a context when someone wants to have represented only the family relationships between persons, or the family and school relations, or any other combination of different dimensions).

5.4 Conclusion

The study and the application of complex networks allowed the representation and visualisation of different real world scenarios by a form of a network. Despite the possibilities addressed by the monodimensional complex network, in order to accurately represent real world situations there is the need to represent more than a single relationship between nodes.

The study on multidimensional analysis permits an approximation closer to the real world scenarios that contains the presence and the distinguish among different type of relations between actors. Either by referring to transportation, social, or utility networks, the multidimensional aspect associated with complex networks allows for a representation more accurate of those scenarios, as well as the possibility to obtain more information related to the type and amount of relationships existent between actors.

This chapter ended with an overview on a study of multidimensional networks that resulted in a new model that considerate each type of multidimensional relationship into single layers. This multilayer network approach allows the isolation of relationships (to serve as a "filter"), or even the creation of different dimensions by combining different types of layer. This approach can result into a network which visualisation is adaptable and configurable depending on different context scenarios or needs.

Chapter 6

Proposed Solution

The previous chapters have addressed the study on personality as a personal characteristic and mediator of our behaviour, associated with the behaviour that is expressed on the usage of Internet platforms. The chapters also noted the growth on the usage and the data present on Internet and its platforms. The demand for data is also linked to the AmI or SmE that seek personal data to enhance their connection to their inhabitants. On a different perspective, a previous chapter also made allusion to the representation of real world scenarios into complex networks either by a monodimensional or a multidimensional approach.

As result, this chapter is dedicated to the definition of the proposed solution that motivated this work. The proposed solution present on this chapter is divided into two separated parts: first part is dedicated to the definition of a virtual OSN based sensor model that intends to provide a wisdom layer to SmE or organisations; the second part is dedicated to the definition of a multidimensional interest model that intends to generate more insights of an individual (or group of individuals) by identifying and classifying different types of interests accordingly to different context needs.

6.1 OSNs As Sensors In Smart Environments

As seen in a previous chapter, both AmI and SmE can benefit knowledge generation regarding their users or inhabitants. The process of knowledge creation is better explained by the Data-Information-Knowledge-Wisdom (DIKW) pyramid, also known as pyramid of knowledge. This concept introduced by Russef L. Ackoff (1989), divides the knowledge creation process into four categories, observable by figure 6.1.

- Data: the lower level of the pyramid. Data is raw, simply exists and has no significance beyond its existence. It can exist in any form, usable or not, and it does not have meaning of itself;
- Information: the process of adding context and meaning to data. Information is data that has been given meaning by a relational connection. This meaning can be useful, but not always does have to be;
- Knowledge: the process of increasing the proper way that information is used. Knowledge is the appropriate collection of information, with intent to be useful. Is a deterministic process, has useful meaning, but it does not provide for, in and of itself, an integration such as would infer further knowledge;

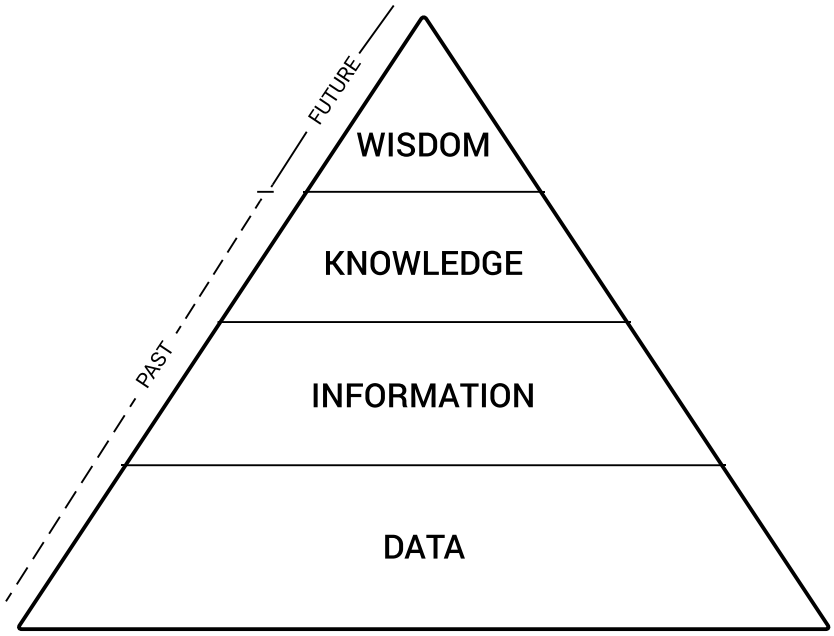


FIGURE 6.1: Representation of the DIKW pyramid, or Pyramid of Knowledge.

- Wisdom: the top level of the pyramid and labelled as a future process. Wisdom is an extrapolation and non-deterministic, non-probabilistic process. It calls upon all the previous levels of consciousness, and specifically upon special types of human programming (moral, ethical codes). It beckons to give us understanding about which there has previously been no understanding, and in doing so, goes far beyond understanding itself. Unlike the previous four levels, it asks questions to which there is no (easily-achievable) answer, and in some cases, to which there can be no humanly-known answer period. Wisdom is therefore, the process by which we also discern, or judge, between right and wrong, good and bad.

The process of knowledge creation begins with data, and due to the undeniable exponential growth of Internet users paired with the consequent growth of OSNs usage, there is a constant enormous quantity and presence of daily data ready to be processed. This fact, among other reasons, motivated the proposed model that is present on this initial part of this work. This model aims to explore the continuous flow of data present in OSNs, with the vision to generate knowledge and provide a wisdom layer to AmI and SmE. Despite the frequent incidence on OSNs, the model is not limited only to that context, in fact, it is intended that the model is applicable to any type of social content present on the Internet.

As result, this work proposes the usage of OSNs as a sensor for SmE. This sensor, represented on figure 6.2 (Barbosa and Santos 2016b), or pseudo-sensor (since usually a sensor associated with psychical devices), has the function of monitoring OSN profiles in order to gather data and generate information and knowledge about a person (or group of persons). By analysing behavioural or network data, that is usually present on every profile, this sensor can provide insights about a persons preferences or behavioural characteristics to the actuators present on the SmE.

This unobtrusive sensor, fulfils the requirements present in the work of Pauwels et al. (2007), and is capable to answer the following questions:

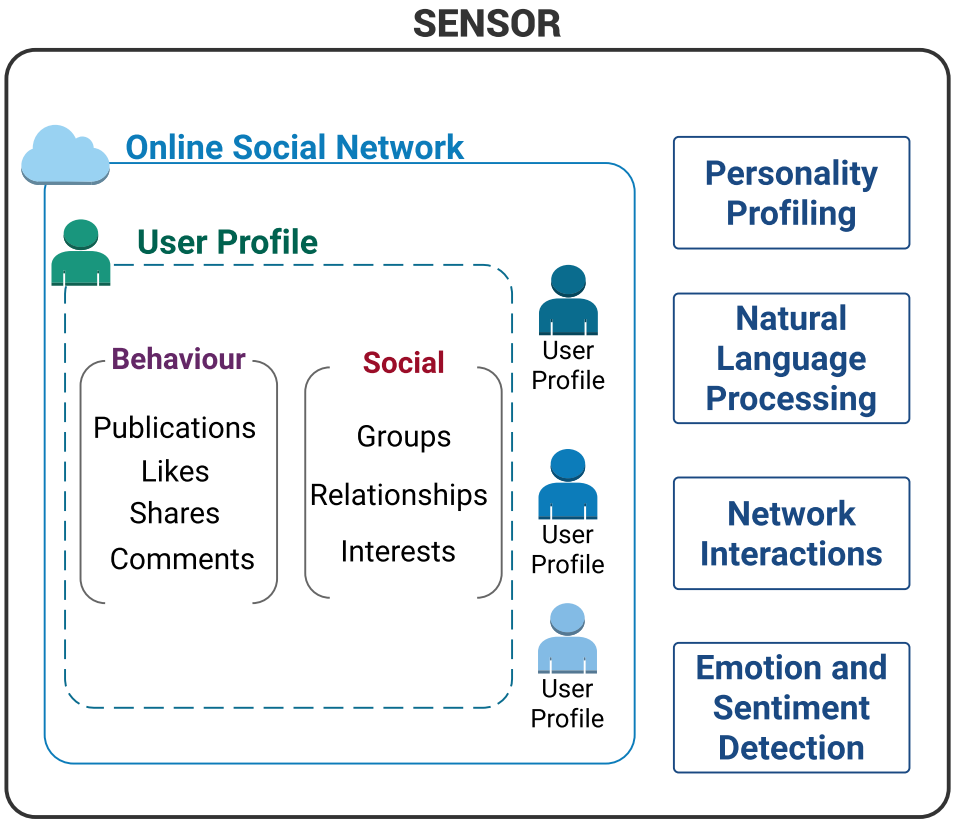


FIGURE 6.2: Virtual OSN based sensor that uses online user behaviour and social references to profile individuals

- Who? by identifying the persons represented/referenced by online social profiles;
- Where and When? by providing a time frame for location, generally present in metadata, in order to determine context;
- What? by recognising activities, interactions, relations, as well as linguistic messages;
- Why? by associating actions with semantics, plans, task identification, and recognising behavioural patterns;
- How? by tracing the information flow trough multiple procedures, recognising behaviour, and literal expressions comprehension.

OSNs offer an immediate significant amount of data ready to be analysed, as well as a history of data by analysing the chronological usage of the platform. Due being a part of the daily life of numerous people around the world, theoretically, there is no need to teach them how to use them on a basic level, which increase the potential and application for this solution.

In addiction to the described characteristics related to OSNs, the focus on this platforms is highly motivated by the work of Kisinski et al. (2013) focused on the accessible digital records of behaviour such as the 'Likes' on the OSN Facebook. The authors proved that on the basis of an average of 68 'Likes' made by an user, it was possible to predict (among other characteristics), the colour tone of their skin (with 95 percent accuracy), sexual orientation (with 88 percent accuracy), or even political affiliation

(with 85 percent accuracy). Other characteristics such as religious affiliation, alcohol, cigarette and drug use, sexual orientation, relationship status, or even if someone parents were divorced are also possible to predict with the same amount of 'Likes' interactions. Kosinski continued his work on the models and, before long, he was able to evaluate any person better than the average work colleague based on 10 'Likes'. When increasing the number of 'Likes' to be taken in consideration, 70 were enough to outdo what the friends of a person knew, 150 what their parents knew, and with 300 what their partner knew. With even more 'Likes' it was possible to even surpass what a person thought they knew about themselves.

What is possible to achieve when considering, in addiction to the 'Likes', the other aspects present on OSN? The proposed sensor model present on this work indents to answer that question by making use of data mining techniques and natural language processing, to extract behavioural clues present in actions (such as publications, 'likes', shares, or comments) that can be used to obtain insights about an individual trough their online profile. In a similar way, network connections like groups association, relationships, or interests, can be used to understand personal preferences (music, movies, brands, events, or even people).

In addiction to the continuous online profile monitoring, this sensor is capable of profiling an individual accordingly to their personality traits, perform natural processing tasks in order to understand context, analyse the different interactions present on their network (identify new relationships, or the addition of new interests), and detect sentiment and emotion present on written language (comments, publications, or other form of written language).

6.1.1 Online Personality Profiling

Currently, OSNs are classified as a good index to predict potential actions of users, with a lot of rich information encoded in the content of those interactions (Moosavi and Jalali 2014). Since, generally, an OSN is shown in a graph and defined as a network of interactions and relationships (where the nodes consist of actors and the edges consist of the relationships or interactions between these actors), is worth analysing the interactions between people and determining structural patterns present on them.

According to Adali and Golbeck (2012), people reveal their personality traits through their use of OSNs, who can be predicted with a relatively high accuracy by analysing public data that people liberally share online. It is possible to say that OSNs are a mirror where users reveal a lot about themselves both in the way they share and how they share-it. In order to understand personality with the scope of online social behaviour, the authors analysed various behaviours of individuals in their social group, and some considered some actions described by the following main groups:

- Network Bandwidth: the amount of overall activity and size of social network, the distribution of activity over time and how long they have been using the OSN;
- Message Content: the type of messages sent, whether they contain URLs (or other types of links) and whether they are forwarded;
- Pair Behaviour: their behaviour towards their friends and followers;
- Reciprocity of actions: to which degree their actions are reciprocated by their friends;

- Informativeness: how informative are various behaviour features across all the friends;
- Homophily: all the previous features computed for the person friends to understand his social circle.

When changing the scope of personality classification into OSNs, Ghavami work (2015) affirms that it is possible to avoid having the standard test scores in order to identify user personality by finding relationships between an user behaviour and personality, or even connection between an user network properties and personality. Based on individual features, some authors (Sulaiman, Rambli and Halim 2011) clustered individuals into four categories :

- Popular Sanguine: Good sense of humour, talkative, enjoy socialising, ability to motivate others, self-centred, disorganised and forgetful of important events;
- Perfect Melancholy: Introvert, analytic, weight everything carefully and thoroughly before making any decision, creative, cautious in making friends, very faithful and compassionate, pay great attention to details, often set high expectations in life, depressed very easily, often view situations negatively and low self-esteem;
- Powerful Choleric: Natural born leader, goal-oriented, strong desire to succeed, confident, charismatic, good at organising tasks, compulsive, often feel the need to be in control of situations and have the tendency to lash out when things are disorganised or people do not follow his instructions;
- Peaceful Phlegmatic: Adaptable to any situations, do not panic easily, remain calm during chaotic situations, very patient, unpredictable, do not express their emotions to others, considered good companion and good listeners, act as mediator in situations of conflict, unenthusiastic, prefer to follow a routine and dislike changes.

Not only it is possible to find behavioural and personality characteristics by analysing the usage of an OSN, but also gender indicators. A study (Moore and McElroy 2012) developed some interesting results where the authors found a significant positive relationship between gender and a number of variables of interest where was possible to find that women spend more time on Facebook, had a greater number of friends, posted more photographs and did more postings about themselves, when compared to men. Although in terms of frequency, women visit their Facebook less frequently than men do.

In terms of personality traits, the study found that more extroverted people have more Facebook friends and report less regret over Facebook content, however, extroversion was not significantly related to time spent, number of photographs or the number of wall postings (either about themselves or others). The results suggest that:

- The high-scorers in agreeableness expressed a greater levels of regret about inappropriate content they may have posted and, surprisingly, they did a greater number of postings about themselves than did the low-scorers;
- Conscientiousness trait was not related to time spent, frequency of use, number of friends or number of photographs, and people with high-score in conscientiousness made significantly fewer wall postings, and expressed more regret than did people with low-score;

- Emotional stability was not significantly related to actual number of friends or photographs, or to the number of wall posting, it was positively related to both how frequently they use Facebook to keep up with others and regret;
- Openness has no significant effect on either Facebook usage or content.

6.1.1.1 Implementation

In order to exemplify the personality classification module, it was chosen a small set of written text produced by a random user on the famous ONS for news and entertainment, Reddit. The choice of this specific platform is associated to the fact that the topics discussed there are all of free opinion and, usually, there is not any severe censorship, which results in opinion more close to the truth from their users. All of these factors are important because, in addition of being a written opinion (in form of text), the content is usually related to the true opinion of their author. This can lead to a higher presence of linguistic style features that can provide information about the authors personality.

As for the topic, due to the natural controversy of the subject, it was chosen a text that expresses the opinion of the user related to the recent United States of America elections. The sample text present on the listing 6.1 represents an excerpt of the data that was used for this analysis. The choice of an English sample of content is justified not only because it is an universal language, but also for the presence of a higher number of scientific content focusing text analysis, as well as being a language with a lack of ambiguities and complicated verbal forms when compared, for example, with Portuguese language.

Therefore, the present text should be able to deduce, with a higher level of confidence value, the personality trait values for this particular user in this particular context.

```
I sincerely hope that's true, particularly if he winds up winning the
    presidency. But I doubt that's the case. Apply Occam's Razor to the
    following two options:
-Every one of Trump's crazy and immature statements since his campaign
    began is an integral part of a master strategy to woo voters who like
    hearing such things. After sealing the Republican nomination he plans to
     pivot hard back to the center and radically alter his behaviour,
    presenting himself as the thoughtful and restrained adult that he
    secretly is.
-Trump is capable of acting like an adult sometimes, when he's in the mood
    for it, and that's enough for him to appear calculating and poised to
    those who believe in him. And sometimes he's not in the mood, so he says
     crazy and immature things.
Option 1 would make him the greatest political tactician in generations.
Option 2 would make him a human being like the rest of us.
I think 2 is far more likely. You see a chessmaster, I see a guy who
    flipped over the chessboard and declared himself the winner, because he'
    s lived a life where he's used to being able to do that.
```

LISTING 6.1: Sample of the text produced by a random Internet user that will be used for classification

Despise the possibility to achieve this results using most of the modern programming languages, for this example it was chosen the open source platform KNIME (KNIME 2016). This platform is written in Java language and it is based on the well known open source multi-language software development, Eclipse. KNIME is focused on data analysis, reporting, and integration and it has been used in pharmaceutical investigation, data analysis, business intelligence, and financial data analysis. The integration with several machine learning and data mining components, as well as the modular concept and the capability to create new modules, make this platform a solid choice for these tasks.

The development process can be divided into several phases, and due to the modular characteristic of the platform (KNIME), this process is facilitated and it becomes easier to define and understand the workflow. Therefore, this process is divided into the following phases:

1. Input definition, tagging, and transformation
2. Data pre-processing
3. Personality traits classification
4. Display of the results

For the first phase of this process, represented by figure 6.3, the first feature required is the capability of reading the desired data. In this case, that task is accomplished by using a document parser module that makes possible the separation and aggregation of all data by person or user. This separation of content allow the classification by context, or perform a full classification by analysing all data available for that specific person.

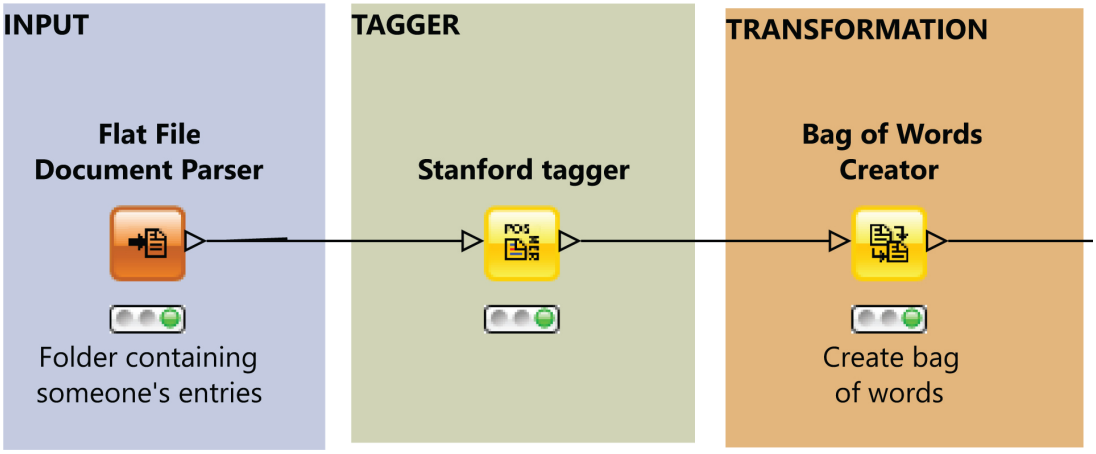


FIGURE 6.3: Input definition, tagging, and transformation phase

After being able to select and import the desired data for analysis, the next step is giving each sentence a set of Part of Speech (POS) tags. This is a common language processing task that automatically assign POS tags to each word in the sentence, such as noun, verb, adjective, and others. Despise the existence of various taggers, for this work it was selected the Stanford Tagger (Stanford 2015), which supports English, Arabic, Chinese, French, and German languages. As for the English tag set, the Stanford Tagger uses the Penn Treebank tag set (Laboratory 2003), and the majority of its tags are represented by table 6.1. This is a vital step because the personality trait attribution will be based on this linguistic characteristics.

TABLE 6.1: Exemplification of the POS tag list used in the Penn Treebank Project

POS TAG	TAG DESCRIPTION
CC	Coordinating conjunction
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plurar
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBN	Verb, past participle

After the attribution of a POS tag to each word of present on the sentences, it is necessary to isolate those words. This leads to the final module of this first phase, the creation of a Bag of Words. This module isolates each word, number, punctuation mark, symbol, or special character found in the sentences. This step eliminates the context meaning, and this is why the input module is so important to isolate content if there is intention to perform a classification based on context.

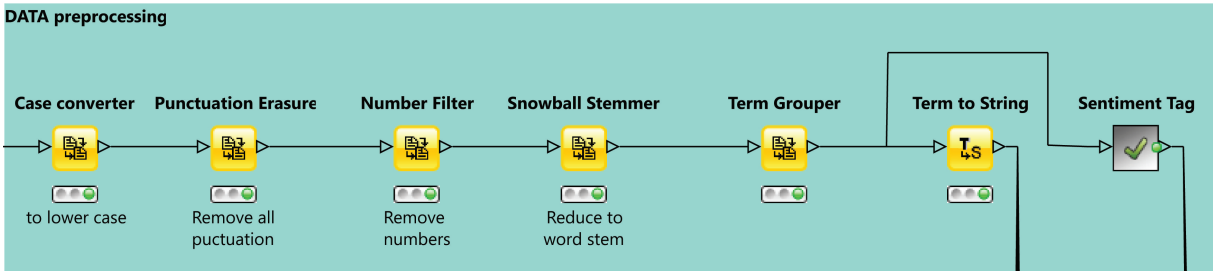


FIGURE 6.4: Data pre-processing phase

With the first phase concluded, it is now time to pre-process all of that data. This phase is responsible to prepare the data in order to be processed by the trait classifiers, and all of the modules required to do so are present on figure 6.4. Making use of the content of the Bag of Words previously created, this input is submitted to the following tasks.

- Case converter: there is the need to uniform the data, so this module is responsible to convert each word to its lower case form;
- Punctuation erasure: punctuation is not going to influence the personality profile result, so this module is responsible to remove each entry that corresponds to punctuation;
- Number filter: similar to the previous module, numbers are not going to influence the personality profile result, and this module is responsible to remove each entry that corresponds to a number;
- Snowball stemmer: this module is responsible to reduce inflected (or sometimes derived) words to their word stem, base or root form (for example, the verb waited is reduced to his stem form which is 'wait');
- Term grouper: the term grouper deletes all of the conflict tags that may exist;
- Term to string: this is a required module because of the data format, until this point the data format is categorised as a 'term' (which includes the words and the tags as a single data entry) and the classifiers modules only accept 'string' format;
- Sentiment tag: this particular module is the result of a combination of modules present on figure 6.5. For this step, each word is crossed against a set of pre-classified positive or negative words (for this pre-classified set it was used the Multi-Prespective Question Answering (MPQA) opinion corpus (MPQA 2015)), in order to identify the sentiment of each word (if a word expresses a sentiment).

The data pre-processing phase output is a group of individual words, in lower case and stem form, classified with a POS tag and a sentiment value (positive or negative). It is now possible to proceed to the personality classification by classifying each trait individually.

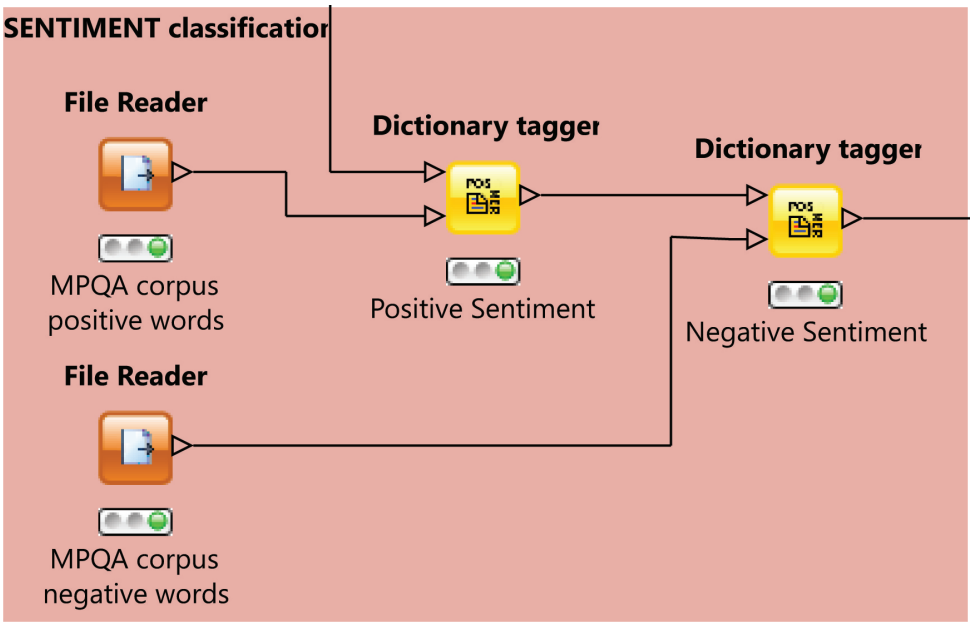


FIGURE 6.5: Sentiment attribution (positive or negative) to each word that match the entries present on the MPQA Corpus data

The process represented on figure 6.6, uses the individual pre-processed data, that is the result of the previous phase, in order to classify each trait based on specific linguistic style features present on table 6.2 (Di Rienzo and Neishabouri 2016). It is important to mention that this personality classification is based on the Big Five model (Goldberg 2006) discussed in a previous chapter.

TABLE 6.2: Linguistic cues for each personality trait

PERSONALITY TRAITS	LINGUISTIC CUES
Openness to Experience	Negative emotions, present verbs, future verbs, third person pronouns, prepositions, articles
Conscientiousness	Positive emotions, present verbs, prepositions
Extroversion	Positive emotions, third person pronouns
Agreeableness	Positive emotions, first person pronouns, present verbs, word longer than 6 letters
Neuroticism	Negative emotions, present verbs, first person pronouns

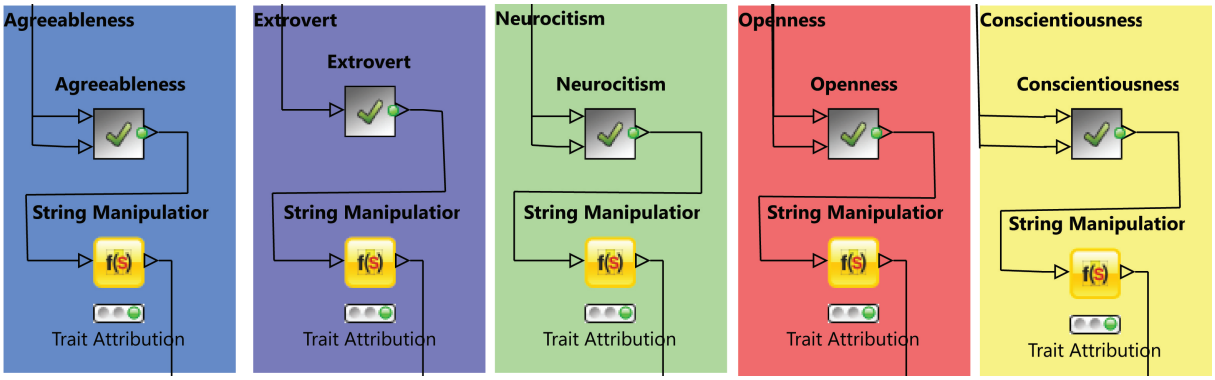


FIGURE 6.6: Personality traits classification phase

After a small task of results preparation, which involves a simple and linear process (grouping the results for each personality trait into a single table), it is possible to observe the results. To each trait was assigned a specific colour, only to facilitate the visual compression of the results, and the result for this specific example is shown in figure 6.7. The figure represents a personality profile based on the text previous discussed and, in this particular case, this profiling is highly biased by context (since the data used belong to the same context). It is possible to achieve more overall results by using data that belongs to different contexts, as well as a higher quantity of data. However, to demonstrate the functionality, the data used served his propose.

By looking at the result in figure 6.7, it is possible to notice a huge predominance of the trait Conscientiousness (41,03%), followed by Openness (27,88%), Agreeableness (19,23%), Neuroticism (8,65%), and finally Extrovert (3,21%). Recalling the characteristics of each trait (Goldberg 2006), and looking at the two highest traits, we have:

- Conscientiousness: responsible, organised, persevering. Conscientious individuals are extremely reliable and tend to be high achievers, hard workers, and planners (organised vs careless);
- Openness to Experience: curious, intelligent, imaginative. High scorers tend to be artistic and sophisticated in taste and appreciate diverse views, ideas, and experiences (insightful vs unimaginative).

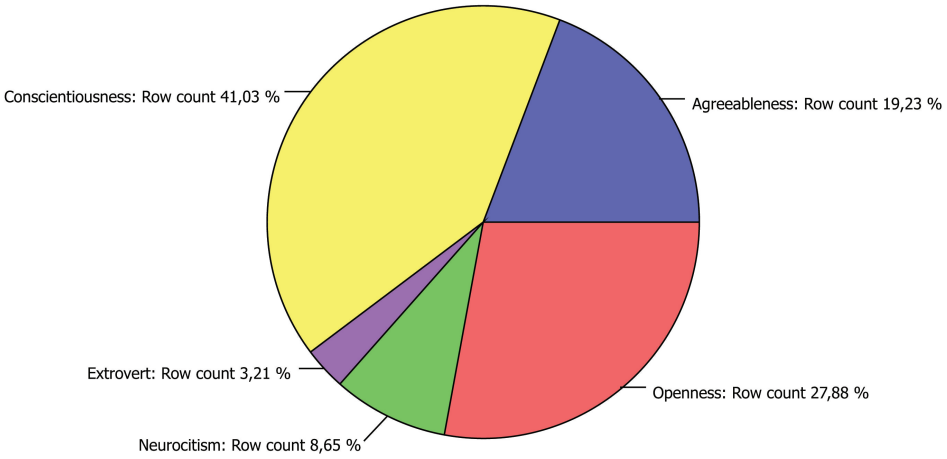


FIGURE 6.7: Result from the personality classification accordingly to the Big Five trait model, using linguistic style features

This results suggests that the author of the text exhibits a tendency to show self-discipline, act dutifully, and aim for achievement. Those are the characteristics of a person that display planned behaviour, and is generally organised. Also the person shows characteristics associated with open mind, and the appreciation of diverse views, ideas, and experiences.

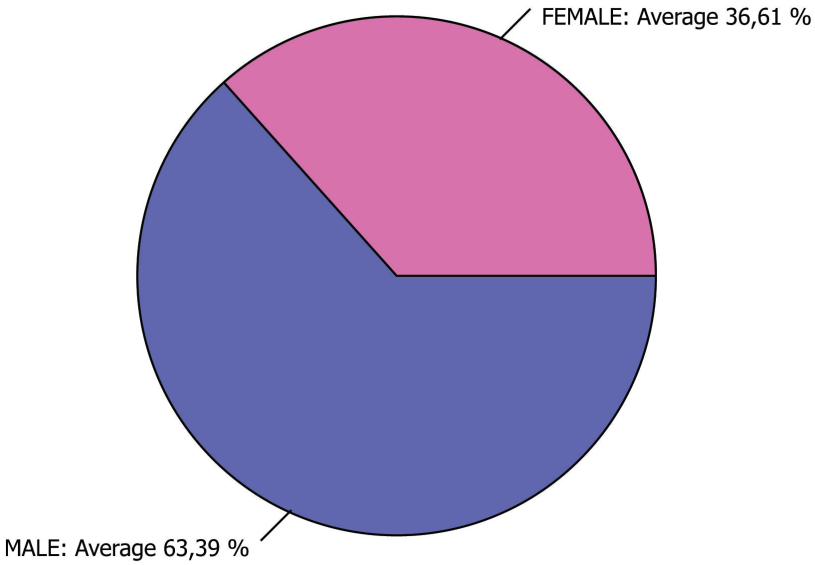


FIGURE 6.8: Result from the gender prediction based on linguistic style features

In addiction to obtaining insights about personality, this process also generated a result for gender prediction by using linguistic style features (in this case the use of pronouns). The result present on figure 6.8, suggests that the author is likely to be a male.

6.1.2 Natural Language Processing

The term Natural Language Processing (NLP) involves a broad set of techniques for automated generation, manipulation and analysis of natural or human languages. Although most NLP techniques inherit largely from linguistics and artificial intelligence, they are also influenced by relatively newer areas such as machine learning, computational statistics, and cognitive science.

Named entity recognition is a sub-task of information extraction that labels sequences of words in a text which are the names of things, such as person, company names, locations or others. In order to understand the meaning of text it is vital to understand and have context recognition, as well as a method for entities identification. One of the most successful solutions for named entity recognition is a result from The Stanford Natural Language Processing Group work (Stanford 2016). Using the sentence present on listing 6.2 as example, it is possible to see in figure 6.9 that the the tagger was able to recognise the different entities that were present in the sentence. Geographical references, organisations, time and date, mathematical references, or even persons identification, are examples of entity references possible to extract from written text.

```
I entered the Amazon.com website and found that smartphone i wanted with a
    discount price. 350 was just the right price i was looking for! It is
    only expected to arrive Monday but Simon said he his 100% sure that he
    will come from London that morning and will get the package for me!
```

LISTING 6.2: Sample of the text that contains reference to entities

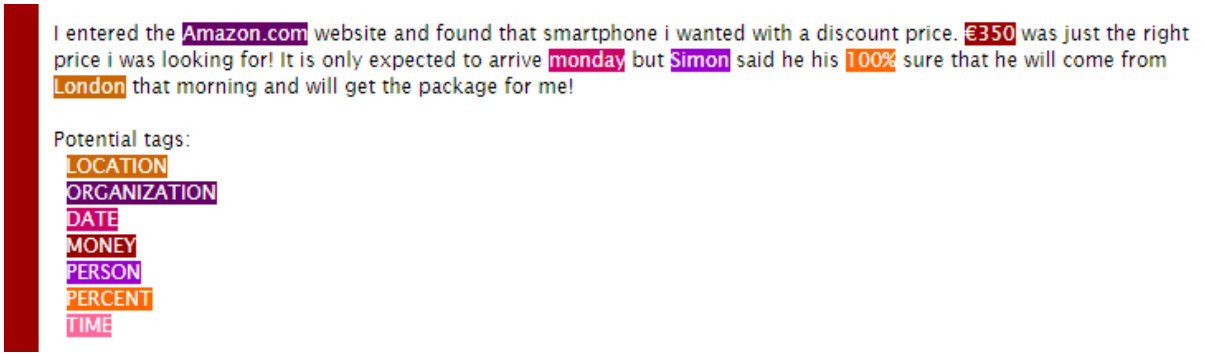


FIGURE 6.9: Named entity recognition using the Stanford tagger

Context is always present on any type of Internet content, and since information is the process of adding context and meaning to data, it is vital that the online virtual sensor is capable of understanding context. This is a vital module that allows the sensor to progress through the DIKW pyramid in order to produce knowledge and, hereafter, wisdom.

6.1.3 Emotion and Sentiment Detection

The process of emotion and sentiment detection is directly associated with the process of sentiment analysis, that consists of the usage of NLP with the intention of deriving sentiment, or subjective information from text. This process is heavily supported by the machine learning concept and their algorithms.

Despite the relation between the terms Artificial Intelligence (AI) and machine learning, and the fact that they become much more widespread than before, in reality they are not quite the same. AI can bee seen as a branch of computer science that attempts to build machines capable of intelligent behaviour, associated with the capability of developing thoughts. In other hand, machine learning is seen as "the science of getting computers to act without being explicitly programmed", it is the implementation of the computer methods that support AI (Bell 2016).

Machine learning evolved from the study of pattern recognition and computational learning, and those characteristics are essential when the topic is emotion and sentiment detection. Machine learning tasks can be classified into the following categories:

- Supervised learning: the common way to solve a learning problem when there are labelled datasets available. In this approach some models are learned from labelled data using learning algorithms and tested against labelled data. This usually yields good results, and it has been widely exploited in personality recognition;
- Unsupervised learning: contrarily to the supervised learning, no labelled dataset is provided to the algorithms. This means that the algorithm is responsible to find structure based on the input;
- Reinforcement learning: a interaction with a dynamic environment in which the computer must achieve a certain goal without knowing when it has come close to it (for example driving a vehicle).

Emotion and sentiment detection may sound familiar since it is a part of the personality profiling module. In fact, the emotion and sentiment detection module works in a similar way, however, due to the necessity of detecting emotion and sentiment present in sentences (or blocks of text) without the personality profiling aspect, the virtual OSN based sensor needs to have a separate dedicated module to that task. The module makes use of supervised learning (similar to the usage of the MPQA opinion corpus), machine learning algorithms like support vector machine combined or probabilistic classifiers like Naive Bayes classifier, in order to fulfil the necessity of detecting emotion and sentiment on textual references.

The choice of an algorithm always depend on the task at hand. While support vector machines tend to perform much better when dealing with multidimensional and continuous features, methods such decision trees or rule based algorithms tend to perform better when dealing with discrete or nominal data, and Naive Bayes method may need a relatively small dataset to achieve the maximum prediction accuracy (when compared to others) (Celli 2013).

6.1.3.1 Naive Bayes Classifier

The Naive Bayes classifier is a supervised statistical and classification method for information classification. The classifier is based on the Bayes Theorem with the "naive" assumption of independence between every pair of features, meaning that the presence (or absence) of a particular feature have no effect or relation with the presence (or absence) of any other feature. This means that the classifier is not looking at entire sentences, but rather at individual words, and this assumption is very strong and super useful since it makes possible to work well with little amounts of data or even data that may be mislabelled.

The Bayes theorem formula is defined by the equation 6.1 (Mohanty 2016).

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \tag{6.1}$$

- $P(c|x)$ is the posterior probability of class (c: target) given predictor (x: attributes).
- $P(c)$ is the prior probability of class.
- $P(x|c)$ is the likelihood which is the probability of predictor given class.
- $P(x)$ is the evidence, the probability of evidence arising without regard for the outcome.

The classifier is particularly useful for very large data sets, along with simplicity, the classifier is known to outperform even highly sophisticated classification methods, and has been mostly used in text classification. Like every method, the classifier has some pros and cons (Ray 2015):

- Is easy and fast to predict class of test data set, as well as multi-class prediction;
- When assumption of independence holds, the classifier performs better compared to other models (like logistic regression), and even needing less training data;
- Performs well with categorical input variables when compared to numeric variables;
- If categorical variable has a category which was not observed in the training data set, the model will assign a zero probability value and becomes unable to make a prediction (often known as "zero frequency");
- The model is also known as a bad estimator, so the probability outputs should not be taken too seriously;
- The assumption of independent predictors might be a limitation since not often is possible to have a set of predictors completely independent.

When searching for applications, this classification model is often seen in the following scenarios:

- Real time Prediction: Naive Bayes is an eager and fast learning classifier, that can be used for real time predictions;
- Text classification/ Spam Filtering/ Sentiment Analysis: This classifier model is often used in text classification (due to better result in multi-class problems and independence rule) have higher success rate when compared to other algorithms. As a result, it is also widely used in Spam filtering (identify spam e-mail) and Sentiment Analysis (associated with social media analysis, with the goal to identify positive and negative customer sentiments);
- Recommendation System: Naive Bayes classifier combined with collaborative filtering are capable to build a recommendation system that uses machine learning and data mining techniques to filter unseen information and able to perform predictions.

In the context of sentiment analysis, the formula can be translated to the formula represented by the equation 6.2.

$$P(sentiment|text) = \frac{P(text|sentiment)P(sentiment)}{P(text)} \tag{6.2}$$

This model is also often seen in the following variations:

- Gaussian: It is used in classification and it assumes that features follow a normal distribution;
- Multinomial: It is used for discrete counts. For example, instead of analyse each word occurring in the document, it counts how often word occurs, and reduces the iterations needed;
- Bernoulli: The binomial model is useful and is at his maximum potential when the feature vectors are binary.

6.1.3.2 Support Vector Machine

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection. This method technique was introduced by Vapnik (1998) and optimised, among others, by Platt (1998). SVMs take the set of training data and marking it as part of a category then predicts whether the test document is a member of an existing class. This methods are often used in text classification, image analysis, or bio-informatics (to name a few).

In order to classify data, the classifier makes use of a hyperplane. A hyperplane is defined as a function which creates a distinct boundary separating two classes, in fact, for a simple classification task with just two features, the hyperplane is actually a line. Another terminology associated with SVMs is called margin. A margin is a distance between the hyperplane and the two closest data points from each respective class (Mohanty 2016).

Similarly to other methods, the SVMs contain some advantages and disadvantages associated to their usage (Scikit 2016):

- SVMs are effective in high dimensional spaces;
- They are still effective even in cases where number of dimensions is greater than the number of samples;
- They make use of a subset of training points in the decision function (called support vectors), so it is memory efficient;
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels;
- However, if the number of features is much greater than the number of samples, the method is likely to give poor performances;
- SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

6.1.3.3 Implementation

As explained before, the virtual OSN based sensor model present on this work, should be able to detect emotion and sentiment present in data in the form of written text. To exemplify the potentials and the application of this module, it was created a simple classifier that aims to fulfil those objectives.

Despise the vast selection of programming languages available today, the sentiment classifier present in this module is written in Python. The Python programming language is a dynamically-typed, object-oriented interpreted language. Although, its primary strength lies in the ease with which it allows a programmer to rapidly prototype a project, a powerful and mature set of standard libraries make it a great fit for large-scale projects as well (Python 2016). Python already has most of the functionality needed to perform simple NLP tasks, however most of the times it does not fulfil the requirements for the most demanding NLP tasks. In order to cover this problem it is required the usage of the Natural Language Toolkit (NLTK) (NLTK 2015). NLTK is a collection of modules and corpora, released under an open-source license, that not only provides convenient functions and wrappers that can be used as building blocks for common NLP tasks, but it also provides raw and pre-processed versions of standard corpora used in NLP literature and courses.

Since a specific programming language is not required, the following implementation description for this module contains a general flow of actions that should be generalist enough to be applied to any programming language or platform of choice (provided that the natural language processing capability is present). The overall vision and function of this module is illustrated by figure 6.10 and this process, when given an input, should generate an output that contains the sentiment classification value of the input (positive, negative, or neutral), as well as a confidence value that indicates a certain value associated with the classification given by the machine learning algorithms. Like referenced previously, in order to obtain an output based on a given input, the module will make use of the NLTK library for the usage of pre-processing tools needed to prepare the data, as well as the well known machine learning classifiers.

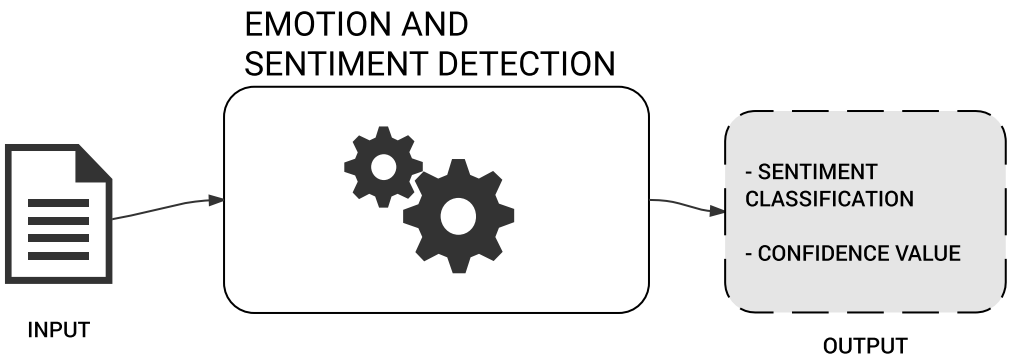


FIGURE 6.10: Overall concept of the emotion and sentiment detection module

Since the module is based on supervised learning, there is a need to define and include the data that will be used to train the classifiers and, in this particular case, is only logical that the data used for training contains sentiment (either positive or negative). In order to achieve this task the module uses two files that contain more than 5000 pre-labelled data (similar to the MPQA opion corpus used on the personality profiling module), positive labelled data (listing 6.3), and negative labelled data (listing 6.4).

```
1 a fascinating and fun film .
  tadpole is a sophisticated , funny and good-natured treat , slight but a
      pleasure .
```

```
3 this insightful , oscar-nominated documentary , in which children on both
      sides of the ever-escalating conflict have their say away from watchful
      parental eyes , gives peace yet another chance .
  i admired this work a lot .
5 whether you're moved and love it , or bored or frustrated by the film , you
      'll still feel something .
   . . . there are enough moments of heartbreaking honesty to keep one glued
      to the screen .
7 my goodness , queen latifah has a lot to offer and she seemed to have no
      problem flaunting her natural gifts . she must have a very strong back .
  a smart , sweet and playful romantic comedy .
9 australian actor/director john polson and award-winning english
      cinematographer giles nuttgens make a terrific effort at disguising the
      obvious with energy and innovation .
```

LISTING 6.3: Pre-labelled positive data sample that will be used for training the classifier

```
  simplistic , silly and tedious .
2 it's so laddish and juvenile , only teenage boys could possibly find it
      funny .
  exploitative and largely devoid of the depth or sophistication that would
      make watching such a graphic treatment of the crimes bearable .
4 [garbus] discards the potential for pathological study , exhuming instead ,
       the skewed melodrama of the circumstantial situation .
  a visually flashy but narratively opaque and emotionally vapid exercise in
      style and mystification .
6 the story is also as unoriginal as they come , already having been recycled
       more times than i'd care to count .
  about the only thing to give the movie points for is bravado -- to take an
      entirely stale concept and push it through the audience's meat grinder
      one more time .
8 not so much farcical as sour .
  unfortunately the story and the actors are served with a hack script .
```

LISTING 6.4: Pre-labelled negative data sample that will be used for training the classifier

After defining the data that will be used for the classifier training process, is needed to process and prepare the data. This process is represented by figure 6.11, where the pre-labelled data will suffer a pre-processing manipulation that consists on the creation of a bag of words with each word, and the correct sentiment tag associated to it. This document classification process extracts each word and grammatical features from sentences, disregarding grammar and order, but keeping multiplicity. Since only words that have impact on the classification are needed for this process, it is possible to discard punctuation or symbols. Each word is then attributed a POS tag, and the process ends with another data transforming task, which converts each word to its lowercase form. This process generates as output a list of individual words, with a POS tag and classified as either positive or negative.

Like every real life scenario, there are words that appear more often than others, and there is the need to exclude some words that appear less frequently, preventing the occurrence of false positives. This task illustrated by figure 6.12 is accomplished by creating a frequency distribution list of all words present

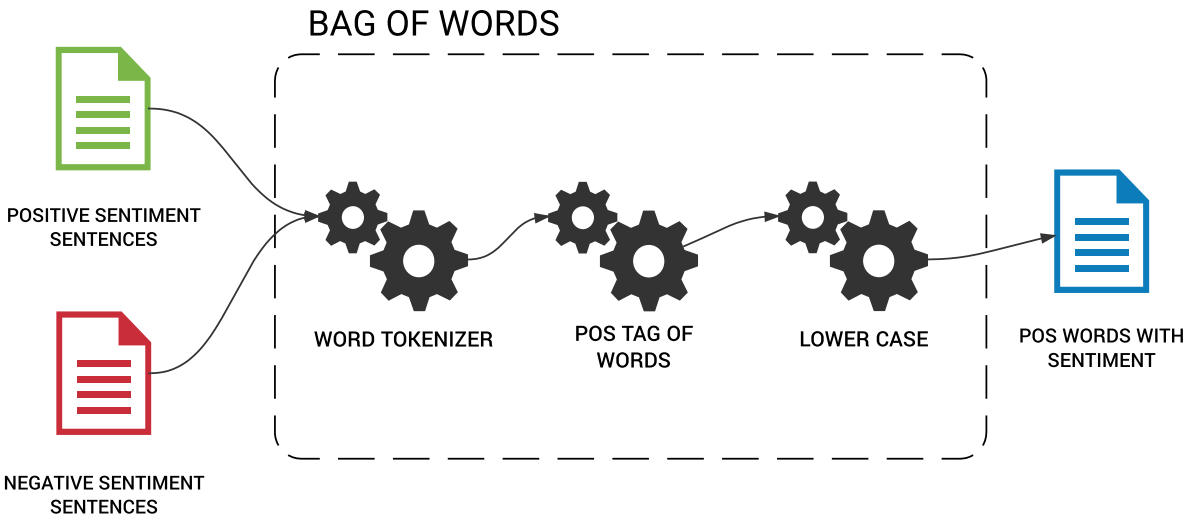


FIGURE 6.11: Inclusion of the pre-labelled data (for positive and negative sentiment) needed for supervised learning

in the dataset that contains words tagged with POS and with a sentiment value associated (positive or negative). This frequency distribution list is order by frequency of occurrence (in a descendant order), meaning that is ordered from the most common word to the least common word.

It would be possible to train our classifiers by using all of this amount of words, however it may not be useful and we should implement a limit. Therefore is needed a filtering process that contains, in this particular example, the 5000 most common words. This filtering process is a crucial step that is directly related to the accuracy value of the classifier, so depending on the dataset used for training, it may need some adjustments and value correction regarding the amount of words to be used. The output of this process is a list containing the top 5000 features (features in NLP vocabulary are known as "keywords" that have a weight on the result (in this case the sentiment value) of a sentence).

FIGURE 6.12: Filtering the dataset using frequency distribution list and filtering the top 5000 most common words

The next step is dedicated to the creation of a training set that will be used to train the classifiers, and a testing set that will be used to measure the accuracy value of the classifiers. Either the testing set and the training set will be based on the top 5000 features list that will be randomised (to prevent the linear

disposition of the features, that contains all positive features followed by all negative features). The training set is composed by the first 2000 entries present on the randomised top 5000 features list, while the testing set is composed by the bottom 2000 entries. The division of data between the training and testing set is intended, and its propose is to prevent biased data .

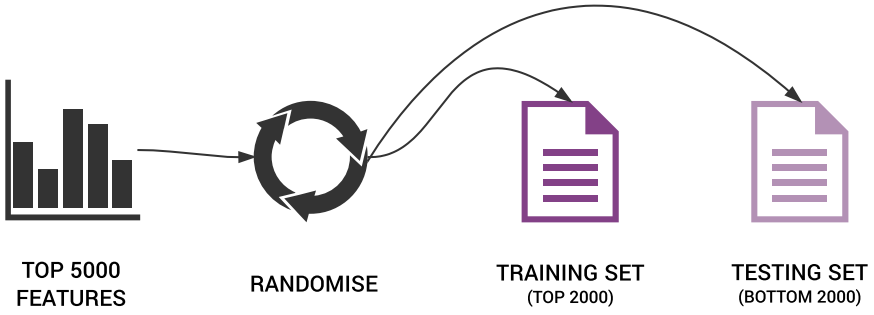


FIGURE 6.13: Defining the training and testing set based on a randomised list of top features

With a training and a testing set defined, is now time to train the classifiers. The classifiers choosen for this module are based on the supervised statistical and classification method Naive Bayes (and variations), and support vector machine methods. The complete list of classifiers used in this emotion and sentiment classification model is the follow:

- Original Naive Bayes classifier: a classifier that follows the statistical and classification method characteristics that were presented in a previous section;
- Multinomial Naive Bayes classifier: is a Naive Bayes classifier for multinomial models. The multinomial Naive Bayes classifier is suitable for classification with discrete features such as word counts for text classification) (ScikitLearn 2017c);
- Bernoulli Naive Bayes classifier: is a Naive Bayes classifier for multivariate Bernoulli models. Similarly to the Multinomial Naive Bayes classifier, this classifier is suitable for discrete data. However, while the Multinomial Naive Bayes classifier works with occurrence counts, the Bernoulli variation is designed for binary/boolean features (ScikitLearn 2017b);
- Linear Stochastic Gradient Descent classifier: simple yet very efficient approach to discriminate learning of linear classifiers under convex loss functions such as (linear) SVMs and Logistic Regression. This classifier has been successfully applied to large-scale and sparse machine learning problems often encountered in text classification and natural language processing (ScikitLearn 2017a);

  The advantages of Stochastic Gradient Descent are:

  - Efficiency;
  - Ease of implementation.

  The disadvantages of Stochastic Gradient Descent are:

  - SGD requires a number of hyperparameters such as the regularisation parameter and the number of iterations;
  - SGD is sensitive to feature scaling.
- Linear Logistic Regression classifier;
- Linear Support Vector Classification classifier: a SVM classifier that uses a linear approach in order to have more flexibility in the choice of penalties and loss functions, and should scale better to large numbers of samples (ScikitLearn 2017d);
- Nu-Support Vector Classification classifier: a SVM classifier where the fit time complexity is more than quadratic with the number of samples which makes it hard to scale to datasets with more than a couple of 10000 samples. A particular characteristic of this classifier is the usage of a parameter that controls the number of support vectors (ScikitLearn 2017e).

The odd number of classifiers is intended because this emotion and sentiment classification module contains a voting system that will determine the classification result of a given sentence. The classifiers training procedure is represented by figure 6.14 where, at first, each classifier is trained based on the contents of the training set. After completing the training phase on each classifier, it is now time to test the accuracy of them. This is accomplished by testing each individual classifier with the testing set that was generated before. In the end, this procedure outputs a list of accuracy values of each classifier. The accuracy values change on each iteration of training and is natural if there are a discrepancy between values between iterations. The accuracy results for the final iteration of this training set that contains 5000 feature words, are represented by the table 6.3.

TABLE 6.3: Accuracy results for each classifier using testing set data with 5000 features

CLASSIFIER	**ACCURACY**
Original Naive Bayes	**82.54%**
Multinomial Naive Bayes	**82.51%**
Bernoulli Naive Bayes	**82.54%**
Linear Stochastic Gradient Descent	**81.71%**
Linear Logistic Regression	**82.22%**
Linear Support Vector Classification	**80.94%**
Nu-Support Vector Classification	**81.86%**

Despise the similarity between accuracy values, is not impossible that classifiers with a similar accuracy value will produce different outputs regarding the sentiment classification of a given sentence (one can determine that the sentence is positive, while the other can determine that is negative). In order to prevent a lack of consensus regarding the classification, in addiction to the odd number of classifiers this module also has implemented a voting system. This voting system uses a statistical implementation of mode. The mode of a sample is the element that occurs with higher frequency in the collection, and since there is present an odd number of classifiers, if, for example, four out of the seven classifiers conclude that one sentence is positive and the remaining three out of seven conclude that is negative, taking in consideration the confidence values of each classifier, is more likely that the sentence is positive. The voting system also provides a solution to the problem when a classifier has a classification with a low confidence value.

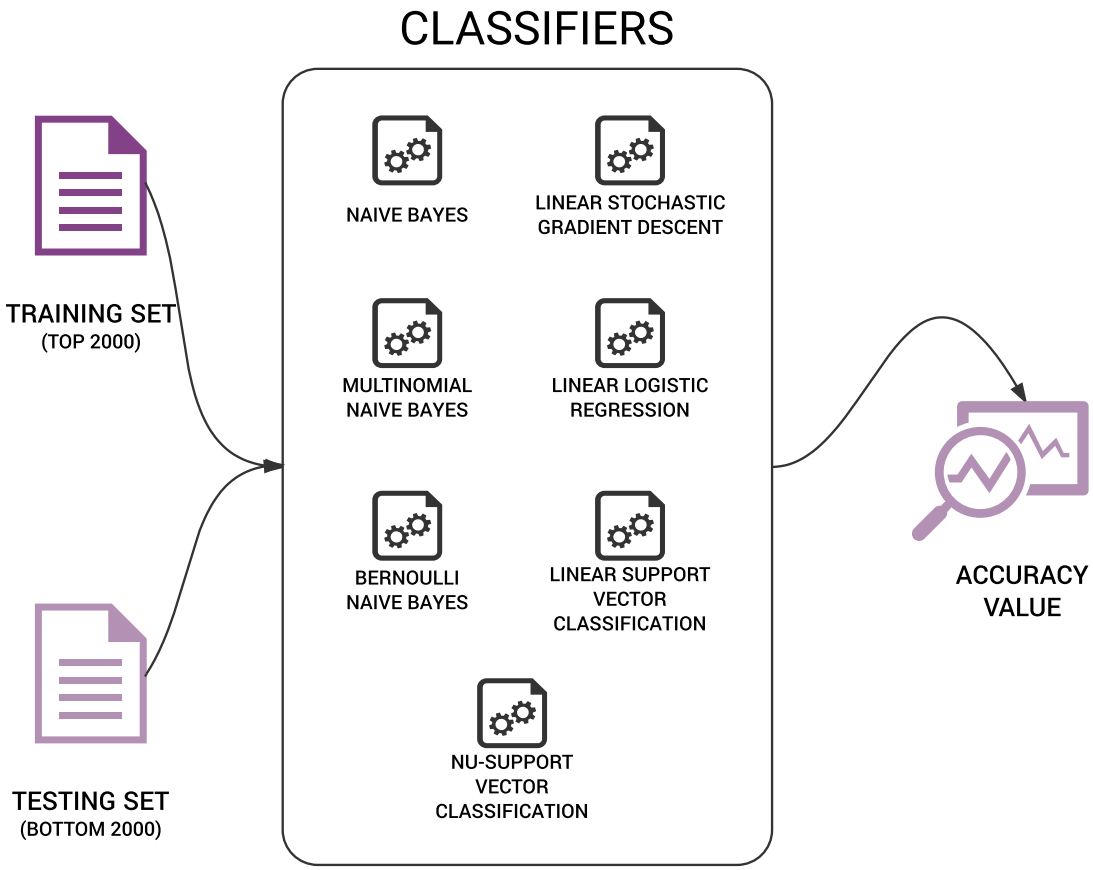


FIGURE 6.14: Training and testing the classifiers. The result of this procedure is an accuracy value for each classifier

The emotion and sentiment detection module is now complete and ready to produce sentiment classification tasks. In order to test and verify that the classification process is working, lets try a test using the following sentence:

- "I really hate it when this happens, just got an awesome TV but the remote does not work"

This sentence will be the input for the module and it is expected that the output will be composed by a classification value and a confidence value. Therefore, the first step to achieve that result is to pre-process the input. This pre-processing phase is similar to the one present in the initial phase of the training set creation, and includes tokenizing words (divide a sentence into single words), attribution of a POS tag, and finally reducing all of the words into their lower case form. In this particular case there is only a sentence that needs classification, but the module would allow a collection of documents as input. This process is described by figure 6.15 results in a list of features that will be then classified by the classifiers previously trained.

With the input pre-processed and with a features list created is now time to classify the sentence mentioned above. The classification process follows a similar guideline when compared to the process of training and testing the accuracy of the classifiers, however, this time the output is not an accuracy value but instead it is expected to be a classification and a confidence value. As observable by figure 6.16, this

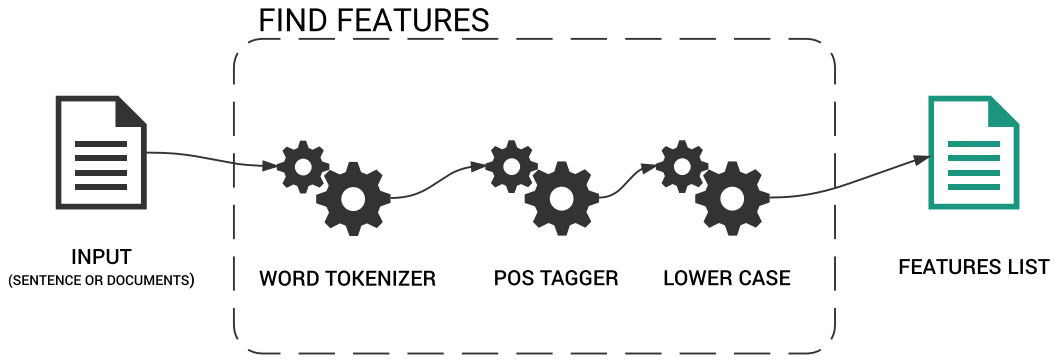


FIGURE 6.15: Pre-processing input to generate a features list ready to be classified

is achieved by running the features list against all of the available classifiers and, after each classifier concludes their classification process, all of the results will be forward to the voting system in order to produce the final output. This final output consists in a combination of a sentiment classification and a confidence value, regarding the input or, in this example, the sentence identified before.

From a human perspective, when looking at the sentence used as the input for this example, it is possible to understand that it contains both ends in terms of sentiment. Is present a clear positive emotion towards the TV, and is in fact described as 'awesome', however the general sentiment of the sentence is negative because the remote that came with the TV does not work, and because this is an event that does not please the owner of the sentence. Therefore, as humans, is safe to presume that the overall sentiment classification for this sentence is negative, but does the emotion and sentiment classification module conclude the same?

```
Input: "I really hate it when this happens, just got an awesome TV but the
    remote does not work"
Result: ('neg', 0.8571428571428571)
```

LISTING 6.5: Output result of the emotion and sentiment detection with sentiment classification and confidence value

The output for the classification of the example sentence is present on listing 6.5 answers the doubt that the emotion and sentiment classification module will perceive the sentence the same way a human does, and the answer is yes. Despise the presence of a positive opinion regarding the TV, the classifier was able to understand the less positive overall situation and classified the sentence as negative with approximately 86% confidence.

6.1.3.3.1 Classifiers Comparison

It is perceivable that each classifier result is different from the others, and is justified by the difference between the algorithms that they are based on, as well as the intention of each classifier. As an example, some classifiers work better with less training data because they were defined with that in mind. This is a natural occurrence that can be seen by the results on table 6.4. This results were obtained by changing the

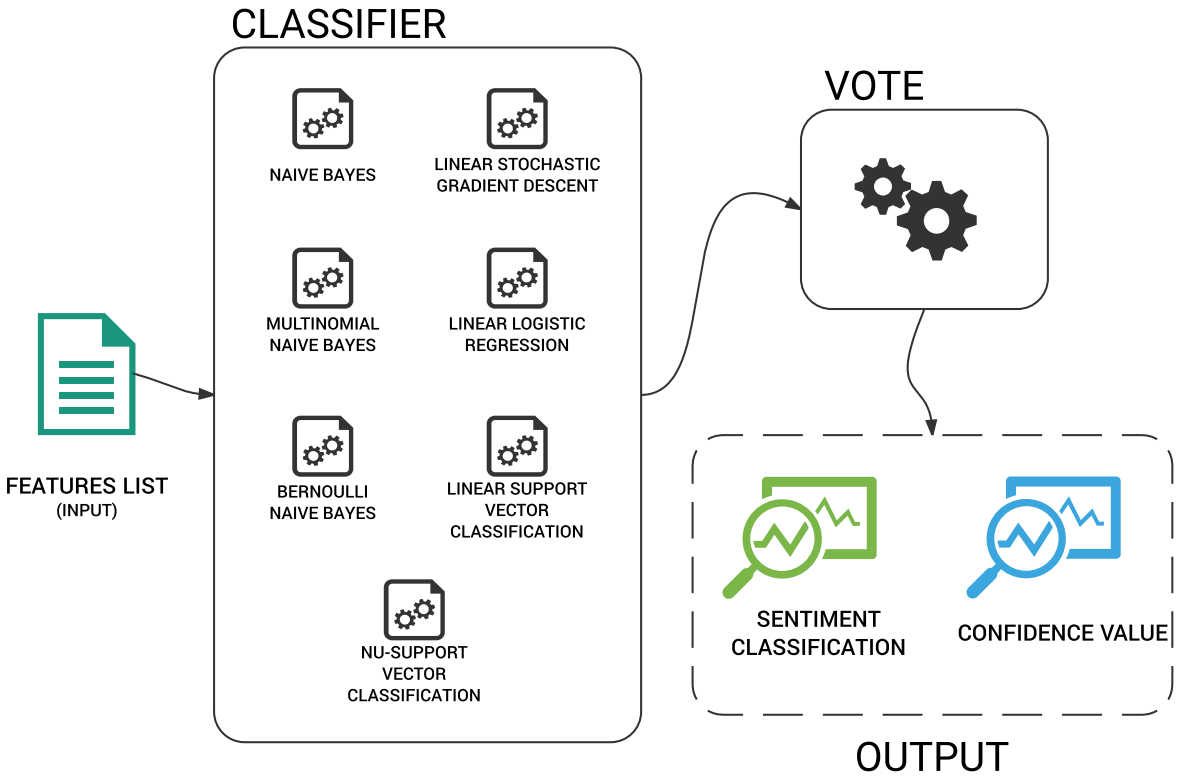


FIGURE 6.16: Classification of a given input, followed by a voting system that will output a sentiment classification and the respective confidence value

size of the training set from a range between 500 and 20000 features. For a better visual perception of the results we have figure 6.17. By looking at the results, it is possible to see that most of the classifiers have a higher accuracy value with the higher training set size, however the Nu-Support Vector Classification and the Naive Bayes Theorem based classifiers have a slight decrease in accuracy at the 20000 features mark.

TABLE 6.4: Classifiers accuracy based on the quantity of features used for training

	Feature Count					
	500	**1000**	**2500**	**5000**	**10000**	**20000**
Original Naive Bayes	77.30%	79.33%	81.44%	82.54%	82.75%	82.44%
Multinomial Naive Bayes	77.28%	79.26%	81.74%	82.51%	82.68%	82.43%
Bernoulli Naive Bayes	77.08%	79.23%	81.50%	82.54%	82.34%	81.37%
Linear Stochastic Gradient Descent	77.25%	79.65%	80.52%	81.71%	82.23%	82.50%
Linear Logistic Regression	78.44%	80.59%	81.38%	82.22%	82.57%	82.78%
Linear Support Vector Classification	78.25%	79.77%	79.44%	80.94%	81.58%	82.05%
Nu-Support Vector Classification	78.23%	80.67%	81.49%	81.86%	82.55%	82.52%

Unless the objective is a really specific task with a limited context, it is not possible to rely solely on one type of algorithm with the expectation that will be the best for every task, it is an unrealistic thought. The combination of different types of algorithms may be the best way to approach a case scenario where the

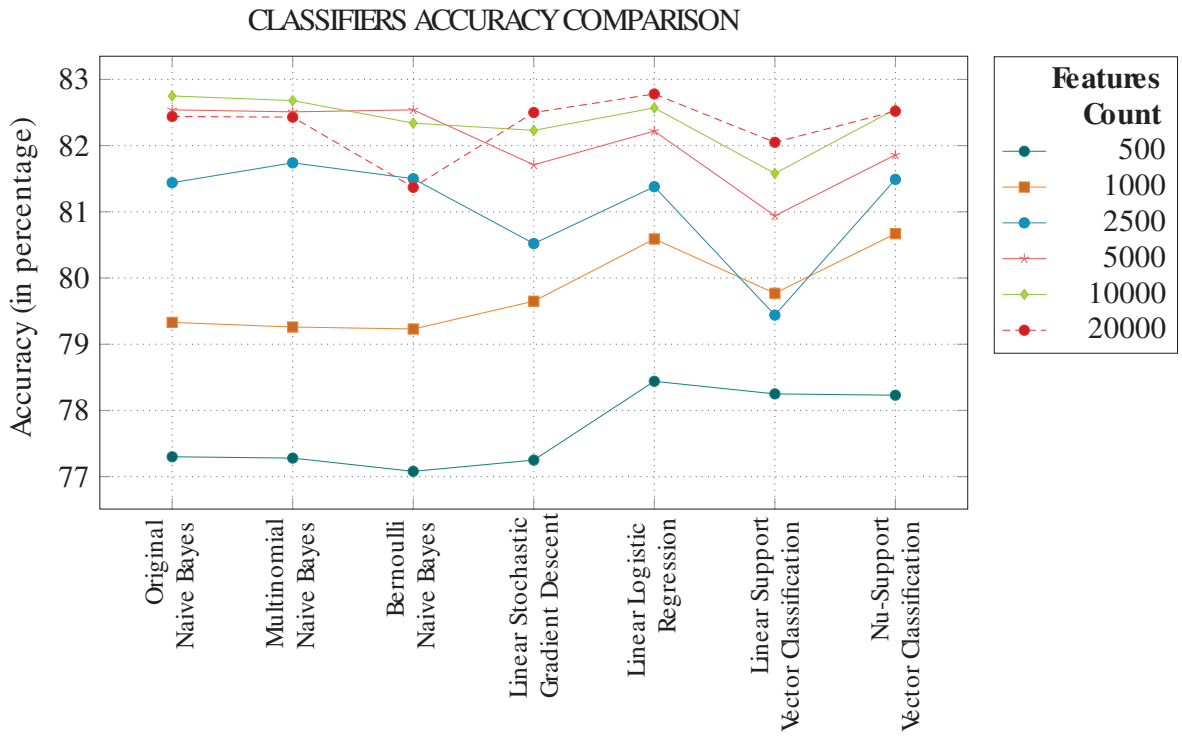


FIGURE 6.17: Comparison of accuracy values from classifiers based on the number of features present on the training set

context borders are not defined and the length of the input is variable. Is important to point that, in either situation, the training process is a vital step and is not discarded the necessity to perform adjustments.

6.2 Multidimensional Interest Network Model

One of the main modules of the virtual OSN based sensor is dedicated the the analysis of OSN interactions. In essence, any OSN is a modelling of a set of nodes (individuals, entities, organisations) and a set of relationships among them (Bouanan et al. 2015), and when given the task of representing an OSN generally it is done by defining a graph with a network of interactions and relationships where the nodes consist of actors and the edges consist of the relationships or interactions between these actors (Moosavi and Jalali 2014). In their core, OSNs are a complex multidimensional network based on the social network aspects of the real world.

The function of this particular module present in this section, is to perform a constant analysis to a given virtual profile (or set of profiles) in order to retrieve or detect changes in the interaction network. This interaction network can be separated into behaviour and social interactions networks, where:

- Behaviour network is based on the expressed online behaviour associated with the following characteristics:
  - Publications;

  - 'Likes';
  - Shares;
  - Comments (or any given type of written based opinion).
- Social interactions network is defined by the social needs that are present on any OSN that are identifiable by the following:
  - Groups;
  - Relationships;
  - Interests.

Despite the effort made in multidimensional networks in the context of OSNs, those works are heavily focused on the social aspect and relations between users (the need to belong). In addiction to the natural heavy focus on social interactions from OSNs, these platforms are also focused on the behavioural actions and the interests of their users, and studies have investigated the relations and similarity between OSN actions ('like', share, comments, and other similar actions) and behavioural intentions (Alhabash et al. 2015; Lee, Ahn and Y. J. Kim 2014).

OSNs already have methods to separate and distinguish different social connections by creating a sub-network containing all connections classified as family, friends, work colleagues, among others. But when given the task of representing and distinguish interests, generally results in a 'monodimensional' network that considerate each different interest as the same and with the same level of personal dedication. Even interests that may appear to be equal, depending on the actions that were performed to identify those interests, it is possible to have two complete different interests in terms of weight value. As an example, an interest produced by clicking the button 'like' may not be the same interest produced by a group affiliation, or an interest expressed by a comment may be different from an interest expressed by a share action (especially since a comment usually requires more time and effort to perform when compared to a single click of a button).

The demand for a better comprehension and representation of individual interests resulted in the second model proposal present on this work, that is focused on the interests that individuals express on the online platform of their choice, or in the general usage of the internet and online services. In order to solve the problems associated with the 'monodimensional' way of representing the interests expressed on OSNs and other online platforms, this work proposes a model for a multidimensional interest network represented by figure 6.18 (Barbosa and Santos 2017b), that can be applied individually or in combination with the virtual online based sensor model (being a part of the network interactions module). By dividing the entire network of interests into individual layers that represent different levels of interest, it is possible to weight each type of action (for example by the effort required to produce each type of action), understand online behaviour, and even understand which type of interests are supported and reinforced by other layers. Based on effort and type of actions, the multidimensional interests network model is divided into three main layers: association, interaction, and opinion (that can be divided into positive and negative opinion).

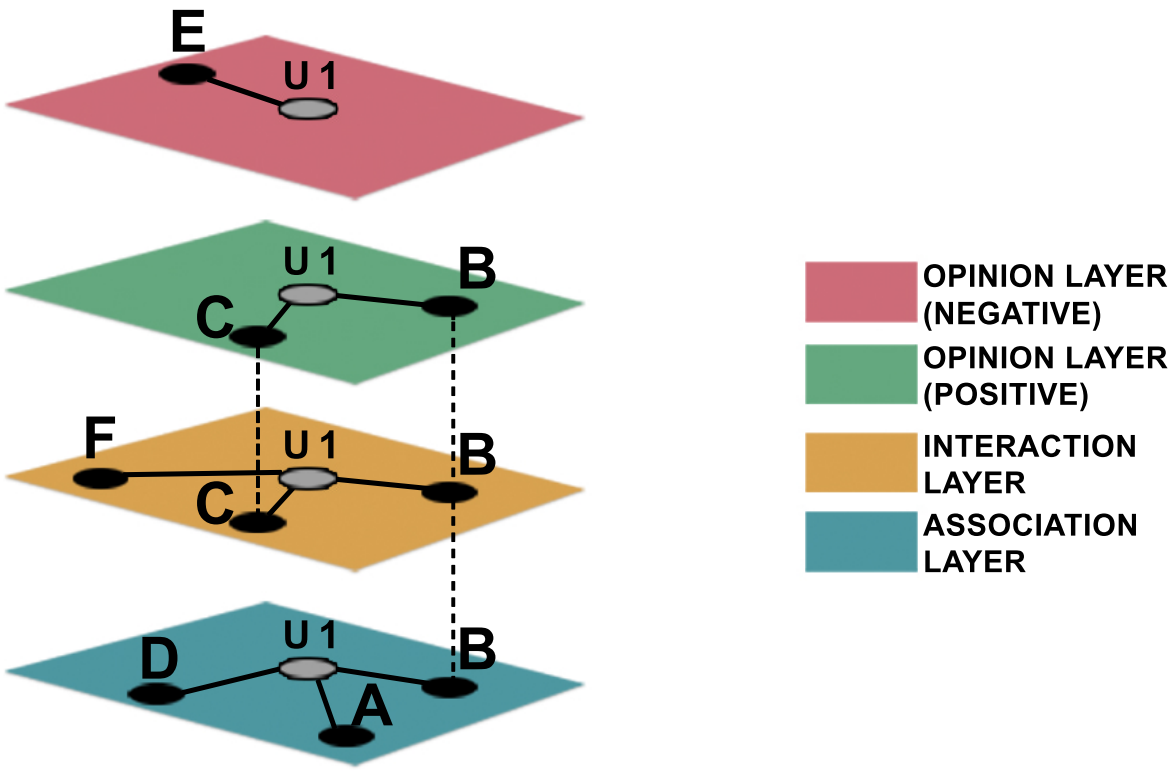


FIGURE 6.18: Multidimensional network representation of an individual interests

In the example provided by the model representation present on figure 6.18 (with opinion layer was divided into two layers, each one containing either positive or negative opinions), in a simple glance is possible to identify a heavily reinforced interest on the node 'B', followed by an interest on node 'C' that is noticeable by the presence of those nodes on multiple layers. If both nodes 'B' and 'C' belong to the same topic (such as music preferences) with this module is it noticeable the appreciation of node 'B' over the node 'D' or even node 'F', that can be translated into a higher level of preference when comparing, for example, music bands. With this type of approach it is also possible to reinforce subjects, for example, if a new node 'D' is created on the interaction Layer, it is automatically reinforced by the presence of the same node on the association layer.

The approach to the interest network provided by this model proposal is also flexible and adaptable when facing the necessity of getting insights about some different levels of interest, or a degree of affiliation, towards a certain topic or subject. By isolating sets of layers, or even combining separate layers, it is possible to create dimensions that fulfil the requirements and the needs of a certain context.

As explained previously, it is possible to define a weight value to each layer depending on the characteristics and the needs of the context. If in a specific context the association and the opinion layer are defined as the most impactful regarding the interests of any given individual, is only natural that the combination of both layers result in the most impactful dimension possible. Therefore, it is possible to isolate those layers and even combining them to create a dimension that better illustrates the interest degree of a specific individual, or group of individuals, regarding a specific context.

The emerge of the necessity to have different dimensions, with different weight values associated with each layer, is exemplified on figure 6.19. This figure illustrates, on left side, a dimension created by the

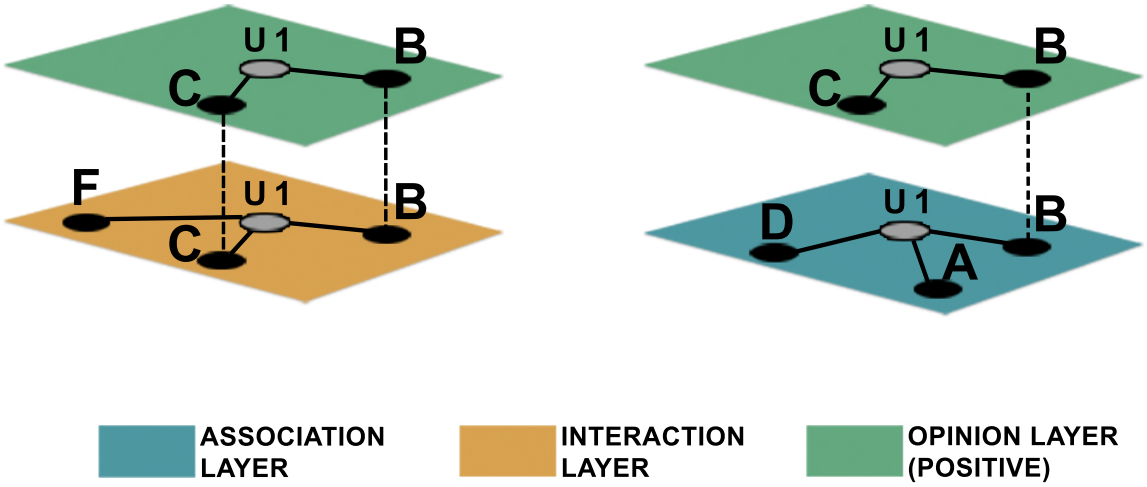


FIGURE 6.19: Multidimensional network representation of an users interest network

combination of the interaction and the positive opinion layers, that are isolated from the general multi-dimensional interest network in order to fulfil the requirements of a specific context. In a similar way, the right side of the figure contains another dimension composed, in this case, by the combination of the association and the positive opinion layer. In this case, is illustrated the capability of isolate and combine different layers even in different rearrangement, however this is not limited to only one individual since the multidimensionality allows a combination of layers that belong to different individuals in order to create a dimension that compare, for example, the interests of two different individuals.

This representation of interests in a multidimensional perspective, associated with the possibility of defining different dimensions composed by different layers, provides another level of insights by displaying and working with data in a way that was not possible by the 'unidimensional' approach.

As stated before, the proposed multidimensional interest network model is composed by three main categories, that result in four layers: association, interaction, positive opinion, and negative opinion. The justification for this arrangement and the characteristics that compose each layer are described bellow.

6.2.1 Association Layer

The bottom layer of the multidimensional interests network model is defined as association layer. This layer is composed by long term relations that can be found on most of modern OSNs, trough actions such as group associations, follows, or even subscriptions. Those actions were paired together due to the fact that they represent a long term relationship commitment that reflects a higher permanent interest on a brand, people, product, or any other type of content, which can be categorised as personal commitment. There is also common characteristics among the actions which constitute this layer and, depending on the OSN platform, there is a semantic characteristic that references one of the described actions by a different name.

In this layer it is expected a constant presence of several nodes and relations and the amount of nodes or relations removals is expected to be low. The justification behind this thought is directly associated to the nature of the actions that compose this layer, since they reflect a higher level of commitment to a specific subject.

6.2.2 Interaction Layer

Following a bottom-top approach to the multidimensional interests model, the next layer is defined as the interaction layer. This particular layer contains a variety of less permanent actions (in terms of commitment to a specific subject), such as 'likes' (or any other similar action that differs only in terms of semantic) and shares (or any other action that serves a sharing purpose). 'Like' and share actions are a fast easy way to share content, they represent appreciation and support for the content (Lee, Ahn and Y. J. Kim 2014), so it is natural that this actions are paired together. These type of actions are considered less permanent due to their ease to perform and lower effort (it is usually a single click).

Either 'likes' or share actions are a result of emotional actions (which have a short duration that may varies from seconds to minutes), and as a result this layer contains a heavy presence of nodes and relations associated with a high frequency of new nodes and relations added. In terms of node and relation removal is highly dependable on the particular behaviour of OSN usage from each individual.

6.2.3 Opinion Layer

The final layer of this model is an opinion layer, that is divided into two layers: positive and negative layer. Overall, this layer represents a network of opinions extracted from written text that is found in comments, publications, or any form of action that uses natural written language. As result, this layer contains a focus on text analysis methods, as well as opinion mining techniques, in order to extract either positive or negative references (in the case of the virtual OSN sensor model this is achieved by the usage of the emotion and sentiment detection module, and the NLP module in order to process and understand context).

Written features present on comments, and any form of publication, provide a dynamic expression of thoughts and feelings with, usually, no restrictions to what is said (Lee, Ahn and Y. J. Kim 2014). Due to the nature of the content present on this layer, it is possible an occurrence of relations that may contradict other relations present on other layers, and this situation is solved with the association of a weight value to each layer (based on the context).

In terms of quantity and frequency of nodes and relations, similar to the interaction layer, is highly influenced by the type of OSN usage from each individual (while some want to produce content, others prefer to passively consume content produced by others).

6.2.4 Exploring The Need For Self-Presentation

People want to self-present themselves in a variety of ways. From verbalisation towards a particular subject or topic (in case of OSN context this is achieved by using written language or the use of other

media such as video), to disclose personal behaviours, self promotion, or beliefs, to surround themselves with associations that reflect their individual preferences, to even construct a public image that represents how they would like to be seen by others.

Virtual platforms, specifically OSNs, allow people to create and promote the image of themselves trough the interactions and demonstrations of interests associated with their OSN usage. The usage of actions such as like, comment, and share can be represented as a way for users to manage their self-presentation by signalling their likes and dislikes, interests, preferences, among other characteristics Lee, Ahn and Y. J. Kim 2014. The exploration of the network of interests can provide information regarding the self-presentation of an individual, and by addressing the network of interests in a multidimensional way, it is possible to obtain more insights about an individual and even fulfill the needs of a specific context.

6.3 Limitations

The usage of any OSN differs from user to user, and since some users may be more active and even want to produce content, in the other end of the spectrum we encounter users that prefer to passively consume content that is produced by others. This is strongly related to the individual personality characteristics of each one of us and results in different online behaviour as well as different amounts of data that is present in their online profiles. Due to these characteristics, the level of insight that is possible to obtain about a certain individual is heavily dependable of the online behaviour of each particular individual.

In addiction the the online behaviour that is influenced by personality characteristics, OSN users are becoming more aware of the information they publish, and more concerned about how this information can be used to identify them. Most concerns related to OSN content is focused on raw demographic information or specific offensively publications. As an example, religious or political affiliation may impact on how an employer views a potential employee, as might an inflammatory post, or even an inappropriate photograph (Wald, Khoshgoftaar and Sumner 2012). In a similar way, and since security is a hot topic nowadays, users are more concerned bout the quantity of personal information that is present on their online profiles.

It is also important to notice that different social media outlets each have different characteristics that will likely affect their effectiveness for personality profiling and, for that reason, it is doubtful that any single personality classifier will provide the best results for all social media. As example, the limit on post length such as on Twitter can lead to unusual grammatical usage which can affect personality profiling trough text analysis (Chin and Wright 2014). Also, even if written language is universal, the social characteristics and the adaptation of modern style features of communication (such the use of visual images like emojis, or the use of slang) can affect the context identification and the profiling process.

6.4 Conclusion

This chapter was divided into two main parts and aggregated the concepts introduced and addressed by previous chapters.

The first part of this chapter was dedicated to the model definition of a virtual OSN based sensor, that is focused on online content, more specifically the content generated by users, in order to process and understand content (by context or emotion), profile each user accordingly to the personality characteristics expressed by the form of written text, and analyse their network interactions that involves group associations, follows, content share, content produced, liked content, among other characteristics. From the Big Five theory, to a more involvement with machine learning techniques, the chapter addresses the different methods and approaches that fulfil the required necessities present on the virtual OSN based sensor, with a presence of small functionality exemplifications.

The second part of this chapter is directly related to the module 'Network Interactions' present on the virtual OSN based sensor but, due to the applicability in other contexts as an individual model, its model definition is present separately. This multidimensional interests network is based on the multidimensional (and sub-consequent multilayer) aspect present on the complex network theory. This model follows the tendencies of the representation of real world scenarios into complex networks, to represent the interests expressed by the usage of online social platforms with the intent to obtain more insights about an individual. This multidimensional approach to interests allows the attribution of different degrees of interests (based on the type of activity that generated that interest) alongside with a multilayer perspective that allows the fulfil of requirements presented by a given context, and the comparison of different types of interests among different types of individuals.

Chapter 7

Theoretical Applications In The Real World

The previous chapter introduced two models focused on the personal characteristics and behaviour that each individual demonstrates by the normal usage of an OSN. The amount of data, and consequent information and knowledge creation produced by the usage of the virtual OSN based sensor and the multidimensional interests network models, intends to provide a wisdom layer to enhance SmE or even to be applied on a organisational perspective. Therefore, this chapter contains some theoretical case scenarios, divided by a SmE or organisational application, with the intent to demonstrate the impact and potential when introducing the virtual OSN based model, the multidimensional interests network model, or a combination of both.

7.1 Organisational Perspective

Alongside with product, services, or content creation, organisations have a constant focus on their customer needs and the relationship that they maintain with them. With the intent to improve customer relationship or even to seek more information related to their customers, organisations have adopted OSN and other social media platforms as a marketing and a communication channel. With these large amount of data that is now presented to them, organisations need to be able to generate knowledge that can help them with decision making, enhance their communication or marketing channels, or even to improve their products.

The following scenarios represent a theoretical application of the virtual OSN based sensor and the multidimensional interests network (in combination or isolated), that can be applied in order to provide the knowledge that organisations seek about their customers and their products, or even enhance the capabilities of an organisation (and is distinguish from the competitors) by providing them information that organisations struggled to obtain and that can help them with decision making processes.

7.1.1 Scenario A

For organisations, the disclosure and presentation of their services or products, trough their marketing department, is a key activity. This activity ensures the smooth and systematic implementation of organisational plans, policies, and other programs that contribute for the control of sale activities, with an

objective to maximise the efficiency and profitability of the organisation. The marketing strategies structures of the organisation are vital to the improvement in customer services and customer satisfaction, and equip organisations with the necessary tools for counteract competition.

In order to fulfil those needs, organisations search for consistent and high quality feedback from their costumers. This can be achieved by asking customers to fill out forms, using feedback sections, or even by reaching out directly to them. Any of these approaches require willingness from the customers to answer any question or to fill any survey, and an extra effort from organisations to obtain the feedback with the consistency and quality they seek.

With organisations following the path of their customers and adopting OSNs (and other social media platforms) as platforms to promote their services and products, new opportunities emerged and, currently, organisations are not taking full advantage of it. The possibility presented for OSN users to express their opinions by the usage of a set of fixed actions (such as 'like'/'dislike' and similar actions that express opinion) and mainly by written language, allied with freedom of speech, allow users to freely express their opinions regarding a specific topic of discussion, a service or product, or even other users or persons.

The virtual OSN sensor, trough the natural language processing and the emotion and sentiment detection modules, can collect and process the written text that online users produce, in order to obtain a constant source of feedback that is, usually, accurate to the opinion of the author that, in this particular case, represents a costumer or a potential costumer. Due to the nature of OSNs architecture, these instances of feedback are also associated with a person and his demographic characteristics that can help organisations to obtain more insights regarding their consumers.

7.1.2 Scenario B

Most organisations contain processes that helps them on customer identification activities, or even to assist the search for potential customers. Those processes are typically based on information directly available to the organisation, normally gathered by asking customers to volunteer information by the completion of a survey, or even by persuading them to join a loyalty program (which customers can decline). Obtaining direct customer information requires an existing relationship with the organisation, meaning that customers need to have either purchased a product or made contact with the organisation in such a way that identification is possible so that additional information can be collected. This is not a reliable method and it means that is not possible to acquire data of unidentified potential customers (Dam and Velden 2015).

The usage of an OSN is associated with the presence of some basic demographic information that is required in order to create a virtual profile. At minimum is required a name, gender and a date of birth, and these fields represent the demographic basis for each virtual profile across various OSNs. In addiction to those basic fields, each user can complete their virtual identity by adding some other information such as home town (or other location reference), contact information, personal interest, job information, and even a profile photograph. The presence of different types of personal data results in different degrees of personal demographic characteristics that is present on each virtual profile.

With demographic information (even in a basic level) is possible to cluster virtual profiles accordingly to demographic characteristics that are dependable on context. The presence of this demographic information can be captured by using the virtual OSN based sensor (associated with the interests of each individual reflected and represented by a multidimensional interests network), and provide organisations with information that can assist the evaluation and understanding of the impact of their content, improve their marketing strategies (by evaluating if the desired content is reaching the desired target audience), or even start targeting their content or their products to specific groups of customers that meet some demographic requirements. In addiction, it provides geographical information that can help organisations to extend (or terminate) their presence in other countries or regions.

7.1.3 Scenario C

In addiction to the demographic characteristics associated with the interests provided by the multidimensional interests network that is represented on other scenario, is possible to add a new layer that can provide more detailed insight about current customers or aid the identification of potential customers. With personality as one of the main characteristics that identify uniquely each one of us, allied with the strong association and influence on behaviour, it represents a deep level of insight about the customers of an organisation, as well as an indicator that can help the prediction of behaviour when facing specific situations or scenarios.

The personality traits identification, provided by the 'personality profiling' module that is present on the virtual OSN based sensor, can provide extra insights by creating an extra layer that can act individually or in combination with the clusters created by demographic characteristics. This personality layer can aid the identification of personality characteristics that are present on a specific cluster of individuals or across multiple clusters. These personality characteristics can reflect behaviour and the presence of a layer that identifies those specific characteristics can help organisations identify, for example, specific individuals or individuals across clusters that possess personality characteristics that are associated with the willingness to try new experiences, or even products.

This results in an approximation to wisdom, that is the top layer present on the DIKW pyramid (or pyramid of knowledge). When facing the presence of a cluster of demographic characteristics that is heavily populated by individuals that contain personality characteristics that makes them more reluctant to change, organisations can anticipate that behaviour when, for example, trying to direct new products to that specific cluster.

In addiction, organisations can anticipate the behaviour of new or potential customers, even if they do not display much data in their interests network. By analysing their individual personality characteristics, along with the demographic characteristics, is possible to anticipate behaviour by analysing the behaviour demonstrated by current customers that possess a similar personality profile and belongs to an interval of demographic characteristics.

7.2 User Perspective

With the increase in popularity and the applications of SmE, originated an increase in the information gathering in order to truly fulfil the requirements and the needs of its inhabitants. With the virtualization of their identities present on OSN or other online social platforms, there is present a large amount of personal data that is publicly available and can enhance the capabilities of any SmE. By using that social media data as a source to generate knowledge and predict the behaviour of their inhabitants, SmE can provide a better environment experience by collecting data on an unobtrusive way. The usage of the virtual OSN based model and the multidimensional interest model can provide that wisdom layer that SmE seek, and the following scenarios describe some examples of its application as well as resolution of current problems that SmE face.

7.2.1 Scenario A

Currently, any type of system or environment that obeys to the concept of AmI is extremely personal. The reason behind this affirmation is the focus on the individualised characteristics of a single individual that leads to a poor capability when addressing the interests and the characteristics of a couple, or a group, of individuals. Excluding the virtual personal assistants (which have a clear focus on a more individualised personal point of view), all of the other scenarios are often populated by more than one individual at a time, and there is a need to fulfil the necessities of a group instead of the necessities of a single individual that is present in that group.

Arguably, deciding which type of music or even specific songs to play for an individual is not a complicated task. Is a simple process of identifying his interests by consulting existing 'playlists', or even history of data related to songs listened by that individual. The difficulty of this task increases exponentially with the amount of individuals to consider in a specific context. Instead of being restrained for a single individual, how to decide which type of music is ideal for two individuals? How about a SmE (for example a room division) that contains a group of people in the context of a party?

The lack of capability from SmE to address these situations is solved by the usage of the models proposed on this work. The virtual OSN sensor is responsible for obtaining insights for each individual by analysing their virtual profiles and identifying key elements that compose each layer of the multidimensional interests network. These key elements are present in a form of written text, group associations, follows, likes, and every other element that is part of a multidimensional layer (association, interaction, opinion (positive/negative)). After that, the multidimensional interests network is responsible for the creation and cross of dimensions (accordingly to the context) in order to fulfil the needs of the environment (in this case the need is finding music that is pleasant for all inhabitants).

As a visual example, figure 7.1 contains a comparison between interests dimensions of two different individuals (can be applied to a group of individuals but for simplicity the example only considers two). If isolating the elements of the layer by context, is possible to have only elements that contain preferences for musics. In terms of layer weights, for this particular context, the negative opinion layer has a higher weight value since playing music that do not please an individual is a scenario to avoid. As possible to observe on the left side of the figure, the dimension composed by the association and interaction layer

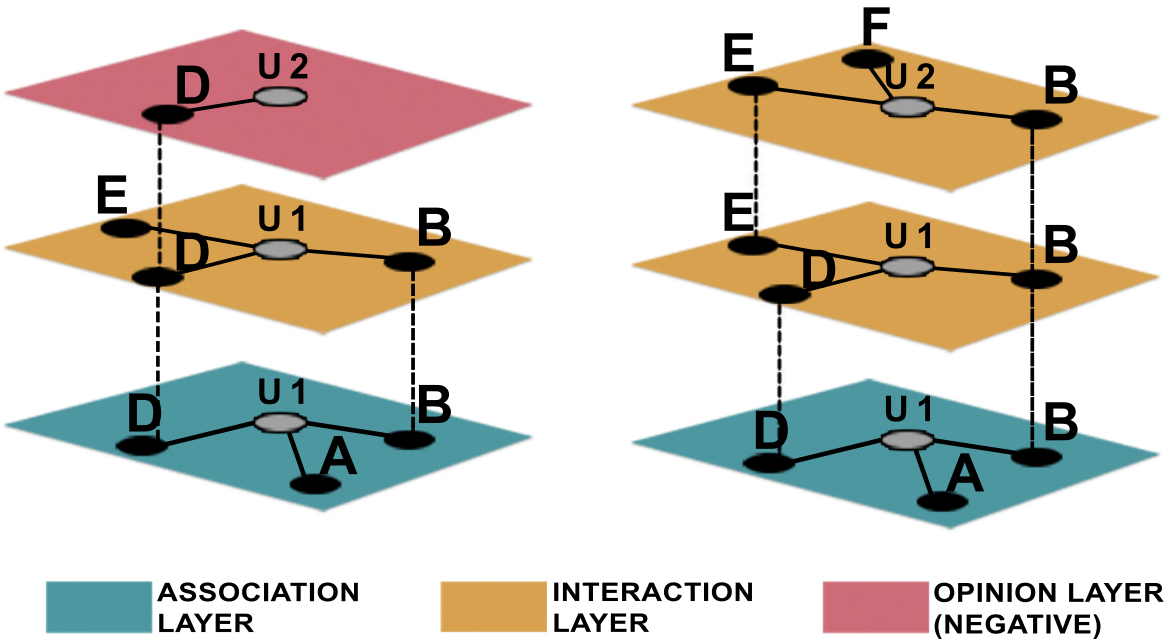


FIGURE 7.1: An analysis between the interests of two individuals using dimensions produced by a multidimensional interests network

that belong to an individual (U1) contain a node 'D' that, despise being an interest of that individual, is a disliked by other individual (U2) that is present on his negative opinion layer. That way the SmE should avoid play any music correlated to the node 'D'. In order to find common elements of interest, the right side of the figure, contains the same U1 dimension present on the left side, but this time is compared against the elements present in the interaction layer of U2. By crossing these dimensions it is found common interests on the node 'E' and 'B', and in terms of weight, the node 'B' is supported by more layers when compared to the node 'E'.

But when is not present enough data about some individual? What is possible to do? This type of problem is addressed by the personality similarities. Considering that our preferences are shaped and influenced by our personality traits, the environment can use history of data about other interest networks and virtual profiles that belong to people with similar personality traits in order to try to predict interests.

7.2.2 Scenario B

Since usually the usage of an OSN is a recurrent activity, is only natural that the interests reflected by associations or interactions are updated or occurs the appearance of new interests. SmE benefit from obtaining insights about their inhabitants and their preferences and interests, and by collecting data related to that interest is a crucial factor to fulfil that objective. By analysing and detecting changes and/or appearance of new interests by association actions (for example group associations, follows, or even subscriptions), by interaction actions (likes, shares, and similar actions), or even by detecting interests present in form of written text. Trough the usage of the natural language processing and the network interactions analysis modules present on the virtual OSN sensor, is possible to detect in a short time any of the occurrences described previously.

If, as an example, an individual shows a new interest on a new music band trough the usage of the interaction action 'like' and the association action 'follow' on the online profile of the respective music band, the virtual OSN sensor recognises that interaction and identifies that new interest. This information can be available to other devices, and when that individual turns on the radio on his car or home, or initiates any type of activity that involves listening to music, the songs from that band are already present on his playlist, as a result of the interest shown on his online social profile.

7.2.3 Scenario C

In a similar way that individuals express their interest trough the usage of OSN, this same usage also contains clues about routines, or even information about activities that break their routine. In a scenario where a SmE has information related to the routines of its inhabitants, the environment benefits from being aware of situations that break routine in order to adjust to a new scenario. Without asking for direct feedback or data to their inhabitants, SmE can adjust even more to their inhabitants needs by constructing knowledge with data that is a result of simple daily activities.

As an example of this particular situation, lets pretend that an individual, trough the usage of his OSN profile, engages on a conversation with a friend or a group of friends about going out for dinner on that night around 9pm. Due to that new schedule activity is less likely that the individual will be in home at that time. The natural language processing module present on the virtual OSN based sensor is responsible to process that written data and understand context in order to 'feed' the SmE with this new routine information. This way, the SmE can adjust its resources to this new scenario by, for example, saving energy by adjusting, or even delaying, the climate control to a later time.

7.3 Conclusion

By an organisational or by a personal perspective (more connected to SmE), both situations require and seek more personal data in order to increase the effectiveness of their tasks and activities, or even to enhance the experience provided to a customer or inhabitant.

Those scenarios can be achieved by the usage of a virtual OSN based sensor and the multidimensional interests network that can aid organisations. Among these scenarios we have customer identification, understanding and measuring the impact of their content, creating a communication channel that includes more a higher frequency and "accurate" feedback from their customers, improving products or services, or even providing an identification method for potential customers or the identification of the needs and interests of new customers.

On a more personal perspective, both models can provide SmE with more personal data that includes interests and patterns of behaviour that can enhance the environment experience around them. Also, the inclusion of the multidimensional interests network, can help SmE to resolve a current problem: the presence of more than one individual in a single environment. With the inclusion of all individuals interests and characteristics, the SmE can direct its actions and shifts on the environment by considering all of the inhabitants instead of being based on the interests of a single one.

Chapter 8

Conclusion and Future Work

8.1 Synthesis

The work present in this dissertation contains two model propositions: a virtual OSN based sensor, and a multidimensional interests network.

Despite the current demand for more personalised solutions,in an organisational or personal perspective, the data that is used to fulfil that goal is often the same. Even with the current adoption of online social platforms as a channel to create and reinforce the relationship between organisations and their customers, organisations are not taking full advantage of the potential that lays in plain sight. The change of focus in the market brought new questions for organisations that besides the answer for the question 'what are we selling?', now need to answer 'Who, What, How, and Why are they buying?'. This new deep level of insight that is required for organisations to improve their marketing strategies, customer relationships, products, or even organisational goals, is also reflected on a personal level on the context of SmE. This environments need to fulfil the necessities of their inhabitants without discarding their preferences and behaviour. Both scenarios have a constant need for personal data and a knowledge or even wisdom creation process associated to it.

By searching for personal data, the motivation and the large amount of individuals that use these online platforms on a daily basis, resulted on a strong and constant presence of personal data that is available directly, or embedded in the context of online interactions. The virtual OSN based sensor model proposed on this work is focused on the content present on OSN (and other online social platforms) in order to identify, analyse, and collect personal data that is present on these platforms. Alongside with the presence of demographic data, there is behavioural and, more specific, personality traits characteristics present in the interactions that each individual performs, or even present on the content of their written text. This model is supported on the study of personality, and the presence of personality traits characteristics on written text, to define a virtual model that, despise not being a physical device, fulfils the requirements defined to achieve a true sensor in a SmE context by being able to identify 'Who', 'Where and when', 'What', 'Why', and 'How'.

The multidimensional interests network model is based on the study of complex networks, more in specific is oriented to the multidimensionality aspect associated with those networks. It is designed to fulfil the necessity of representing and classifying different types of interests. Until now, the interests demonstrated by individuals (either on real world scenarios or by the usage of OSN or other online

platforms) is often displayed with the same level of importance (or degree). This is not an accurate approach because, like in the real world, our interests have different degrees of importance or impact. Therefore, the implementation of this model contains an attribution of degrees to each type of interest in order to fulfil the requirements of a given context. The model also contains a multilayer characteristic that allows the creation of different dimensions with different layers of interests, or even to perform comparisons between the interests of different individuals. This is a direct answer to the necessity of SmE to consider the different types of interests of different types of individuals that are present on the same environment, but also can be applied on other scenarios.

The implementation of the models proposed in this dissertation (on a combined or isolated perspective), represents an immediate impact on organisations or SmE. Customer identification, obtaining direct and "accurate" feedback, identification of needs and preferences, identification of potential customers and their needs, definition of a true SmE that considers the interests of all of is inhabitants, and enhancement of the capabilities of a SmE, are only the representation of a few scenarios of application. While the full range of possible scenarios is only limited by our imagination.

8.2 Scientific Contribution

During the elaboration and development of this work some scientific contributions were made, namely:

- "Online social networks as sensors in smart environments" presented on Smart and Secure Environments track of the 2016 Global Information Infrastructure and Networking Symposium (GIIS) (Barbosa and Santos 2016b). This contribution identifies the explored association with personality and written language, and the correlation with behaviour, associated with the growth and presence of content in OSNs, to present the virtual OSN based sensor model.
- "Layer by Layer: A Multidimensional Approach to Online Social Profiles" in publication and to be presented on the Social Computing session of the Computing Conference 2017 (formerly called Science and Information (SAI) Conference) (Barbosa and Santos 2017a). This contribution explores the presence of interests on OSN and proposes a multidimensional interests network model to represent those interests that introduces a way to compare, cross, and classify interests on various degrees depending on the desired context.
- "Multidimensional Approach to Online Interest Networks" published in DEStech Transactions on Computer Science and Engineering (Barbosa and Santos 2017b). This contribution explores an organisational view and scenarios that are possible by the usage and combination of the virtual OSN based sensor together with the multidimensional interests network.
- "Multidimensional Interests Trough An Online Social Network Sensor For Smart Environments" published in Journal of Information Systems & Operations Management (Barbosa and Santos 2016a). This contribution explores the scenarios, inclined to a individual and to a smart environment perspective, by associating the virtual OSN based sensor with the multidimensional interests network.

8.3 Future Work

Despite the current contributions present on this work this cannot be considered a "completed work" since it can be improved in terms of capabilities and by reduce and resolving the limitations identified.

- Since personality can be highly related to verbal or written language expression, and the personality traits are influenced by the language or idiom, is possible that an individual can express two or more different personality traits depending on the idiom used for expressing himself. Despite the current support for the idiom and the lexical characteristics manifested in English language, is required an extension on the idiom applications that, for geographical and idiom characteristics of the author of this work, will be addressed on a first phase to the Portuguese language and its lexical characteristics. Other idioms will be implemented on future phases by "popularity" order.
- This current work is focused on OSNs by the reasons that were described on previous chapters, however, since other Internet platforms are populated by data, the virtual OSN based sensor should extend is capabilities to other platforms (such as discussion forums, or any given platform that allows the liberal expression of opinion).
- In addiction to written language, OSNs and other Internet platforms are also populated by other types of media (pictures, videos, or even sounds). These type of media can contain additional data that can contribute for the enrichment of the knowledge produced by the virtual OSN based sensor and the multidimensional interests network. Deep learning techniques focused on image recognition can aid the detection of brands or other types of interests that are present on pictures in order to produce a new multidimensional interest node, or even enrich the existent ones. The addiction of this feature results in the creation of a new multidimensional layer that is related to all of the interests that are gathered from other types of media besides written text.
- Even by addressing content by lexical features, it may contain the presence of words or other features that can invert the meaning of the sentence (like double negation or even the use of sarcasm). In addiction, the written text that is produced today (in the context of social interactions) often contains visual expressions (like smiles or emojis) that can enrich or change the meaning of the sentence. Despite the popularity on computer vision and speech recognition fields, deep learning for natural language processing has proved to be successful for sentence classification purpose and can provide a new layer of comprehension about a specific content or context in order to help understand key features that can have an extra impact on a sentence or content.
- Either in a organisational context or a personal context associated with SmE, there is the need of a simulation environment with agents, that can help to understand how the environment reacts and responds to the information that is given to it. Also, this simulation environment will be used to understand the impact of the content produced by a virtual profile of an organisation, and by simulating the virtual profiles of their customers, evaluate the reactions and the behaviour produced when facing certain types of content. This can help organisations to improve their virtual images and the impact of their content.

Bibliography

Ackoff, R.L. (1989). 'From Data to Wisdom'. In: *Journal of Applied Systems Analysis* 16, pp. 3–9.

Adali, S. and J. Golbeck (2012). 'Predicting Personality with Social Behavior'. In: *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE, pp. 302–309. ISBN: 978-1-4673-2497-7. DOI: 10.1109/asonam.2012.58. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6425747.

Agarwal, Basant (2014). 'Personality Detection from Text : A Review'. In: *International Journal of Computer System* 1.1, pp. 1–4.

Ajzen, I. (2005). *Attitudes, Personality And Behaviour*. Mapping social psychology. McGraw-Hill Education, ISBN: 9780335224005. URL: http://www.ebook.de/de/product/21806843/i_ajzen_attitudes_personality_and_behaviour.html.

Alhabash, Saleem et al. (2015). 'From Clicks to Behaviors: The Mediating Effect of Intentions to Like, Share, and Comment on the Relationship Between Message Evaluations and Offline Behavioral Intentions'. In: *Journal of Interactive Advertising* 15.2, pp. 82–96. ISSN: 1525-2019. DOI: 10.1080/15252019.2015.1071677.

Amichai-Hamburger, Yair and Gideon Vinitzky (2010). 'Social network use and personality'. In: *Computers in Human Behavior* 26.6, pp. 1289–1295. ISSN: 07475632. DOI: 10.1016/j.chb.2010.03.018.

Antunes, Mario, Diogo Gomes and Rui Aguiar (2013). 'Towards behaviour inference in smart environments'. In: *2013 Conference on Future Internet Communications (CFIC)*. IEEE, pp. 1–7. ISBN: 9781479900596. DOI: 10.1109/cfic.2013.6566324.

Argamon, Shlomo et al. (2005). 'Lexical Predictors of Personality Type'. In: *Proceedings of joint annual meeting of the interface and The Classification Society of North America* 50.1, pp. 21–21. ISSN: 0002-9572. DOI: 10.2105/ajph.50.1.21.

Bachrach, Yoram et al. (2012). 'Personality and patterns of Facebook usage'. In: *Proceedings of the 3rd Annual ACM Web Science Conference on - WebSci '12*. ACM Press, pp. 24–32. ISBN: 9781450312288. DOI: 10.1145/2380718.2380722. URL: http://dl.acm.org/citation.cfm?doid=2380718.2380722.

Barbosa, Ricardo and Ricardo Santos (2016a). 'Multidimensional Interests Trough An Online Social Network Sensor For Smart Environments'. In: *Romanian Economic Business Review* 10.2, pp. 299–310. URL: https://ideas.repec.org/a/rau/journl/v10y2016i2p299-310.html.

– (2016b). 'Online social networks as sensors in smart environments'. In: *2016 Global Information Infrastructure and Networking Symposium (GIIS)*. IEEE, pp. 1–6. DOI: 10.1109/giis.2016.7814950.

– (2017a). 'Layer by Layer: A Multidimensional Approach to Online Social Profiles'. In:

– (2017b). 'Multidimensional Approach to Online Interest Networks'. In: *DEStech Transactions on Computer Science and Engineering* cmee, pp. 1–5. ISSN: 2475-8841. DOI: 10.12783/dtcse/

cmee2016/5372. URL: http://dpi-proceedings.com/index.php/dtcse/article/view/5372.

Bell, Lee (2016). *Machine learning versus AI: what's the difference?* URL: http://www.wired.co.uk/article/machine-learning-ai-explained (visited on 26/11/2016).

Berlingerio, Michele et al. (2013). 'Multidimensional networks: Foundations of structural analysis'. In: *World Wide Web* 16.5–6, pp. 567–593. ISSN: 1386145X. DOI: 10.1007/s11280-012-0190-4.

Bouanan, Youssef et al. (2015). 'Modeling and Simulation of Human Reaction in a Multidimensional Social Network'. In: *IFAC Proceedings Volumes (IFAC-PapersOnline)* 48.3, pp. 592–597. ISSN: 14746670. DOI: 10.1016/j.ifacol.2015.06.146.

Bródka, Piotr and Przemyslaw Kazienko (2014). 'Multilayered Social Networks'. In: *Encyclopedia of Social Network Analysis and Mining*. Ed. by Reda Alhajj and Jon Rokne. New York, NY: Springer New York, pp. 998–1013. ISBN: 978-1-4614-6170-8. DOI: 10.1007/978-1-4614-6170-8_239. URL: http://dx.doi.org/10.1007/978-1-4614-6170-8%5C_239.

Bródka, Piotr, Przemysław Kazienko et al. (2012). 'Analysis of Neighbourhoods in Multi-layered Dynamic Social Networks'. In: *International Journal of Computational Intelligence Systems* 5.3, pp. 582–596. DOI: 10.1080/18756891.2012.696922. eprint: http://dx.doi.org/10.1080/18756891.2012.696922.

Celli, Fabio (2013). *Adaptive Personality Recognition from Text*. LAP Lambert Academic Publishing, p. 124. 124 pp. ISBN: 365935404X. URL: http://www.ebook.de/de/product/20375194/fabio_celli_adaptive_personality_recognition_from_text.html.

Chaffey, Dave (2016). *Global Social Media Research Summary 2016*. URL: http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/ (visited on 17/11/2016).

Chin, D.N. and W.R. Wright (2014). 'Social media sources for personality profiling'. In: 1181, pp. 79–85.

Contractor, Noshir, Peter Monge and Paul Leonardi (2011). 'Network Theory | Multidimensional Networks and the Dynamics of Sociomateriality: Bringing Technology Inside the Network'. In: *International Journal of Communication* 5. ISSN: 1932-8036. URL: http://ijoc.org/index.php/ijoc/article/view/1131.

Cook, Diane J. and Wenzhan Song (2009). 'Ambient Intelligence and Wearable Computing: Sensors on the Body, in the Home, and Beyond'. In: *J. Ambient Intell. Smart Environ.* 1.2, pp. 83–86. ISSN: 1876-1364. URL: http://dl.acm.org/citation.cfm?id=1735835.1735836.

Cottone, Pietro, Gabriele Maida and Marco Morana (2014). 'User Activity Recognition via Kinect in an Ambient Intelligence Scenario'. In: *IERI Procedia* 7, pp. 49–54. ISSN: 22126678. DOI: 10.1016/j.ieri.2014.08.009. URL: http://www.sciencedirect.com/science/article/pii/S2212667814000288.

Cottone, Pietro, Giuseppe Lo Re et al. (2013). 'Motion sensors for activity recognition in an ambient-intelligence scenario'. In: *2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*. March. IEEE, pp. 646–651. ISBN: 9781467350778. DOI: 10.1109/percomw.2013.6529573.

Cristani, Matteo, Erisa Karafili and Claudio Tomazzoli (2015). 'Improving Energy Saving Techniques by Ambient Intelligence Scheduling'. In: *2015 IEEE 29th International Conference on Advanced Information Networking and Applications*. Vol. 2015-April. IEEE, pp. 324–331. ISBN: 9781479979042. DOI: 10.1109/aina.2015.202.

Dam, Jan-Willem van and Michel van de Velden (2015). 'Online profiling and clustering of Facebook users'. In: *Decision Support Systems* 70, pp. 60–72. ISSN: 01679236. DOI: 10.1016/j.dss.2014.12.001.

Di Rienzo, Antonella and Asana Neishabouri (2016). 'Recommendations with personality traits extracted from text reviews'. In: *Studies in Computational Intelligence*. Vol. 616, pp. 355–364. DOI: 10.1007/978-3-319-25017-5_33.

Ding, Ke et al. (2015). 'Towards building a word similarity dictionary for personality bias classification of phishing email contents'. In: *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*. IEEE, pp. 252–259. DOI: 10.1109/icosc.2015.7050815.

Ducatel, K et al. (2001). 'ISTAG Scenarios for Ambient Intelligence in 2010'. In: *Society*, p. 58. URL: ftp://ftp.cordis.europa.eu/pub/ist/docs/istagscenarios2010.pdf.

eMarketer (2016). *Nearly One-Third of the World Will Use Social Networks Regularly This Year*. URL: https://www.emarketer.com/Article/Nearly-One-Third-of-World-Will-Use-Social-Networks-Regularly-This-Year/1014157 (visited on 15/11/2016).

Forestier, Mathilde, Julien Velcin and Djamel Zighed (2011). 'Extracting Social Networks to Understand Interaction'. In: *2011 International Conference on Advances in Social Networks Analysis and Mining*. IEEE, pp. 213–219. ISBN: 9780769543758. DOI: 10.1109/asonam.2011.64.

Ghavami, Seyed Morteza et al. (2015). 'Facebook user's like behavior can reveal personality'. In: *2015 7th Conference on Information and Knowledge Technology (IKT)*. IEEE, pp. 1–3. DOI: 10.1109/ikt.2015.7288797.

Golbeck, Jennifer et al. (2011). 'Predicting Personality from Twitter'. In: *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*. IEEE, pp. 149–156. ISBN: 9780769545783. DOI: 10.1109/passat/socialcom.2011.33.

Goldberg, Lewis R. (2006). 'An alternative "description of Personality": the Big-Five factor structure.' In: *Journal of personality and social psychology* 59.6, pp. 1216–1229.

Grossman, Gary (2016). *The next stop on the road to revolution is ambient intelligence*. URL: https://techcrunch.com/2016/05/07/the-next-stop-on-the-road-to-revolution-is-ambient-intelligence/ (visited on 15/11/2016).

Gupta, Umang and Niladri Chatterjee (2013). 'Personality Traits Identification Using Rough Sets Based Machine Learning'. In: *2013 International Symposium on Computational and Business Intelligence*. IEEE, pp. 182–185. ISBN: 978-0-7695-5066-4. DOI: 10.1109/iscbi.2013.44. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6724349.

Higgins, E Tory (2000). 'Does personality provide unique explanations for behaviour? Personality as cross-person variability in general principles'. In: *European Journal of Personality* 14.5, pp. 391–406. ISSN: 0890-2070 1099-0984. DOI: 10.1002/1099-0984(200009/10)14:5<391::aid-per394>3.0.co;2-6. URL: http://doi.wiley.com/10.1002/1099-0984(200009/10)14:5%3C391::AID-PER394%3E3.0.CO;2-6.

InternetLiveStats (2017). *Number of Internet Users (2016)*. URL: http://www.internetlivestats.com/internet-users/ (visited on 17/04/2017).

Kim, Jongkwang and Thomas Wilhelm (2008). 'What is a complex graph?' In: *Physica A Statistical Mechanics and its Applications* 387.11, pp. 2637–2652. DOI: 10.1016/j.physa.2008.01.015.

Kivela, M. et al. (2014). 'Multilayer networks'. In: *Journal of Complex Networks* 2.3, pp. 203–271. ISSN: 20511329. DOI: 10.1093/comnet/cnu016.

KNIME (2016). *KNIME Open Source Story*. URL: https://www.knime.org/knime-open-source-story (visited on 25/11/2016).

Kosinski, M., D. Stillwell and T. Graepel (2013). 'Private traits and attributes are predictable from digital records of human behavior'. In: *Proceedings of the National Academy of Sciences* 110.15, pp. 5802–5805. DOI: 10.1073/pnas.1218772110. eprint: http://www.pnas.org/content/110/15/5802.full.pdf. URL: http://www.pnas.org/content/110/15/5802.abstract.

Laboratory, LINC (2003). *Alphabetical list of part-of-speech tags used in the Penn Treebank Project*. URL: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html (visited on 24/11/2016).

Lee, Eunsun, Jungsun Ahn and Yeo Jung Kim (2014). 'Personality traits and self-presentation at Facebook'. In: *Personality and Individual Differences* 69, pp. 162–167. ISSN: 01918869. DOI: 10.1016/j.paid.2014.05.020.

Magnani, Matteo, Anna Monreale et al. (2013). 'On multidimensional network measures'. In: *Italian Conference on Sistemi Evoluti per le Basi di Dati (SEBD)*, pp. 1–8.

Magnani, Matteo and Luca Rossi (2011). 'The ML-Model for Multi-layer Social Networks'. In: *2011 International Conference on Advances in Social Networks Analysis and Mining*. IEEE, pp. 5–12. ISBN: 9780769543758. DOI: 10.1109/asonam.2011.114.

Marceau/Peyrouse (2009). *Slam ma muse*. Sebastiani 2002. CORNAC. ISBN: 9782895291305. DOI: 10.1145/1461928.1461959. URL: https://www.amazon.com/Slam-ma-muse-Marceau-Peyrouse/dp/2895291306?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=2895291306.

Markovikj, Dejan et al. (2013). 'Mining Facebook data for predictive personality modeling'. In: *Proceedings of the 7th international AAAI conference on Weblogs and Social Media (ICWSM 2013), Boston, MA, USA*, pp. 23–26.

McGrath, Felim (2015). *Top 10 Reasons for Using Social Media*. URL: http://www.globalwebindex.net/blog/top-10-reasons-for-using-social-media (visited on 17/11/2016).

Mekhalfi, Mohamed L. et al. (2016). 'Recovering the sight to blind people in indoor environments with smart technologies'. In: *Expert Systems with Applications* 46, pp. 129–138. ISSN: 09574174. DOI: 10.1016/j.eswa.2015.09.054.

Mohanty, Soumendra (2016). *Explaining Machine Learning to a 5th Grader*. URL: http://www.iamwire.com/2016/09/explaining-machine-learning-to-a-5th-grader/141231 (visited on 26/11/2016).

Moore, Kelly and James C. McElroy (2012). 'The influence of personality on Facebook usage, wall postings, and regret'. In: *Computers in Human Behavior* 28.1, pp. 267–274. ISSN: 07475632. DOI: 10.1016/j.chb.2011.09.009. URL: http://linkinghub.elsevier.com/retrieve/pii/S0747563211002020.

Moosavi, Seyed Ahmad and Mehrdad Jalali (2014). 'Community detection in online social networks using actions of users'. In: *2014 Iranian Conference on Intelligent Systems (ICIS)*. IEEE, pp. 1–7. ISBN: 9781479933518. DOI: 10.1109/iraniancis.2014.6802552.

MPQA (2015). *MPQA Opinion Corpus Release Page*. URL: http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/ (visited on 25/11/2016).

Nadkarni, Ashwini and Stefan G. Hofmann (2012). 'Why do people use Facebook?' In: *Personality and individual differences - Elsevier* 52.3, pp. 243–249. ISSN: 0191-8869. DOI: 10.1016/j.paid.2011.11.007. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3335399&tool=pmcentrez&rendertype=abstract.

Newman, M E J (2003). 'The structure and function of complex networks'. In: *E-Print Cond-Mat/0303516* 45.2, pp. 167–256. URL: http://arxiv.org/abs/cond-mat/0303516.

NLTK (2015). *Natural Language Toolkit*. URL: http://www.nltk.org/ (visited on 26/11/2016).

Oberlander, Jon and Alastair J. Gill (2006). 'Language With Character: A Stratified Corpus Comparison of Individual Differences in E-Mail Communication'. In: *Discourse Processes* 42.3, pp. 239–270. ISSN: 0163-853X. DOI: 10.1207/s15326950dp4203_1.

Ortigosa, Alvaro, Jose Ignacio Quiroga and Rosa M. Carro (2011). 'Inferring user personality in social networks: A case study in Facebook'. In: *2011 11th International Conference on Intelligent Systems Design and Applications*. IEEE, pp. 563–568. ISBN: 978-1-4577-1675-1. DOI: 10.1109/isda.2011.6121715.

Paunonen, Sampo V. (2003). 'Big Five factors of personality and replicated predictions of behavior.' In: *Journal of Personality and Social Psychology* 84.2, pp. 411–424. ISSN: 0022-3514. DOI: 10.1037/0022-3514.84.2.411.

Pauwels, Eric J., Albert A. Salah and Romain Tavenard (2007). 'Sensor Networks for Ambient Intelligence'. In: *2007 IEEE 9th Workshop on Multimedia Signal Processing*. IEEE, pp. 13–16. ISBN: 978-1-4244-1273-0. DOI: 10.1109/mmsp.2007.4412806. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4412806.

Pennebaker, James W. (2011). 'Your Use of Pronouns Reveals Your Personality'. In: *Harvard Business Review*.

Pennebaker, James W. and Laura A. King (1999). 'Linguistic styles: Language use as an individual difference.' In: *Journal of Personality and Social Psychology* 77.6 (6), pp. 1296–1312. DOI: 10.1037/0022-3514.77.6.1296.

Pitcher, Rod (2014). 'Revealing the Colour and Personality in Texts : Putting the " Person " Back into our Results'. In: 19, pp. 1–8.

Platt, John C. (1998). *Advances in Kernel Methods: Support Vector Learning*. Ed. by Bernhard Schölkopf, Christopher J. C. Burges and Alexander J. Smola. Cambridge, MA, USA: The MIT Press. Chap. Fast Training of Support Vector Machines Using Sequential Minimal Optimization, pp. 185–208. ISBN: 0-262-19416-3. URL: https://www.amazon.com/Advances-Kernel-Methods-Support-Learning/dp/0262194163?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0262194163.

Poria, S.a b et al. (2013). *Advances in Soft Computing and Its Applications*. Vol. 8266 LNAI. PART 2. Springer, pp. 484–496. 560 pp. ISBN: 3642451101. DOI: 10.1007/978-3-642-45111-9_42. URL: http://www.ebook.de/de/product/21648178/advances_in_soft_computing_and_its_applications.html.

Python (2016). *About Python*. URL: https://www.python.org/about/ (visited on 26/11/2016).

Raisinghani, Mahesh et al. (2006). 'Ambient Intelligence: Changing Forms of Human-Computer Interaction and their Social Implications'. In: *Journal of Digital Information* 5.4. ISSN: 1368-7506. URL: https://journals.tdl.org/jodi/index.php/jodi/article/view/149.

Ramos, Carlos, Juan Carlos Augusto and Daniel Shapiro (2008). 'Ambient Intelligence - the Next Step for Artificial Intelligence'. In: *Progress in Artificial Intelligence* 23.2, pp. 15–18. ISSN: 03029743. DOI: 10.1109/mis.2008.19. URL: http://dl.acm.org/citation.cfm?id=1782254.1782282.

Ray, Sunil (2015). *6 Easy Steps to Learn Naive Bayes Algorithm*. URL: https://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/ (visited on 26/11/2016).

Revelle, William and Klaus R Scherer (2005). 'Personality and emotion'. In: *Handbook of personality and affective science*, pp. 304–305.

Santoro, Richard (2016). *2016 Global Social Media Research Summary For Business*. URL: http://socialmedia-authority.com/2016/08/21/2016-global-social-media-research-summary-business/ (visited on 17/11/2016).

Scikit (2016). *Support Vector Machines*. URL: http://scikit-learn.org/stable/modules/svm.html (visited on 26/11/2016).

ScikitLearn (2017a). *1.5. Stochastic Gradient Descent scikit learn 0.18.1 documentation*. URL: http://scikit-learn.org/stable/modules/sgd.html (visited on 27/04/2017).

– (2017b). *sklearn.naive_bayes.BernoulliNB scikit learn 0.18.1 documentation*. URL: http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html (visited on 27/04/2017).

– (2017c). *sklearn.naive_bayes.MultinomialNB scikit learn 0.18.1 documentation*. URL: http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html (visited on 27/04/2017).

– (2017d). *sklearn.svm.LinearSVC scikit learn 0.18.1 documentation*. URL: http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html#sklearn.svm.LinearSVC (visited on 27/04/2017).

– (2017e). *sklearn.svm.NuSVC scikit learn 0.18.1 documentation*. URL: http://scikit-learn.org/stable/modules/generated/sklearn.svm.NuSVC.html#sklearn.svm.NuSVC (visited on 27/04/2017).

Seidman, Gwendolyn (2013). 'Self-presentation and belonging on Facebook: How personality influences social media use and motivations'. In: *Personality and Individual Differences* 54.3, pp. 402–407.

Selterman, Dylan (2012). *The "Need to Belong" - Part of What Makes Us Human*. URL: http://www.scienceofrelationships.com/home/2012/4/16/the-need-to-belong-part-of-what-makes-us-human.html (visited on 15/11/2016).

Shannon Greenwood, Andrew Perrin and Maeve Duggan (2016). *Demographics of Social Media Users | Pew Research Center*. URL: http://www.pewinternet.org/2016/11/11/social-media-update-2016/ (visited on 17/11/2016).

Smith, Kit (2016). *96 Amazing Social Media Statistics and Facts for 2016*. URL: https://www.brandwatch.com/blog/96-amazing-social-media-statistics-and-facts-for-2016/ (visited on 17/11/2016).

Socievole, A. et al. (2015). 'ML-SOR: Message routing using multi-layer social networks in opportunistic communications'. In: *Computer Networks* 81, pp. 201–219. ISSN: 13891286. DOI: 10.1016/j.comnet.2015.02.016.

Stanford (2015). *Stanford Log-linear Part-Of-Speech Tagger*. URL: http://nlp.stanford.edu/software/tagger.shtml (visited on 25/11/2016).

– (2016). *Stanford Named Entity Recognizer*. URL: http://nlp.stanford.edu/software/CRF-NER.shtml (visited on 26/11/2016).

Statista (2017a). *Most famous social network sites worldwide as of January 2017, ranked by number of active users (in millions))*. URL: http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (visited on 27/03/2017).

– (2017b). *Number of social media users worldwide from 2010 to 2020 (in billions)*. URL: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ (visited on 20/02/2017).

Stavropoulos, Thanos G. et al. (2015). 'Rule-based approaches for energy savings in an ambient intelligence environment'. In: *Pervasive and Mobile Computing* 19, pp. 1–23. ISSN: 15741192. DOI: 10.1016/j.pmcj.2014.05.001.

Stevenson, Angus (2010). *Oxford Dictionary of English*. URL: //www.oxfordreference.com/10.1093/acref/9780199571123.001.0001/acref-9780199571123.

Sulaiman, S., D. R. A. Rambli and Khairul Amirah Abdul Halim (2011). 'A study on the relationship between personality traits and image perceived'. In: *2011 IEEE International Symposium on IT in Medicine and Education*. Vol. 2, pp. 419–423. DOI: 10.1109/ITiME.2011.6132138.

Udo-Imeh, Philip Thomas et al. (2015). 'Personality and Consumer Behaviour: A Review'. In: *European Journal of Business and ManagementOnline)* 7.18, pp. 2222–2839.

Vapnik, Vladimir N. (1998). *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer. ISBN: 0-387-94559-8. URL: https://www.amazon.com/Nature-Statistical-Learning-Theory/dp/0387945598?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0387945598.

Wald, Randall, Taghi Khoshgoftaar and Chris Sumner (2012). 'Machine prediction of personality from Facebook profiles'. In: *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)*. IEEE, pp. 109–115. ISBN: 9781467322843. DOI: 10.1109/iri.2012.6302998.

Weiser, M., R. Gold and J. S. Brown (1999). 'The origins of ubiquitous computing research at PARC in the late 1980s'. In: *IBM Systems Journal* 38.4, pp. 693–696. ISSN: 00188670. DOI: 10.1147/sj.384.0693. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5387055%5Cnhttp://ieeexplore.ieee.org/xpls/abs%5C_all.jsp?arnumber=5387055.

Zephoria (2016). *The Top 20 Valuable Facebook Statistics – Updated November 2016*. URL: https://zephoria.com/top-15-valuable-facebook-statistics/ (visited on 15/11/2016).

ESCOLA
SUPERIOR
DE TECNOLOGIA
E GESTÃO
POLITÉCNICO
DO PORTO

P.PORTO

M

MESTRADO
Engenharia Informática

Individuals Recognition and Simulation Based on Multiple Data Sources

Ricardo Manuel Pacheco Barbosa