Towards the Study of Human Emotions Through Social Media Contents

Manuel Barreto Lima Reis

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisor: Prof. Bruno Emanuel da Graça Martins

Examination Committee

Chairperson:	Prof. José Manuel da Costa Alves Marques
Supervisor:	Prof. Bruno Emanuel da Graça Martins
Member of the Committee:	Prof. Fernando Manuel Marques Batista

November 2015

Abstract

Emotions are present in the everyday life of human-beings, and are therefore reflected in the way humans communicate and interact with each other. Gestures, prosody, and the words we choose among a given context, are also good reflectors of the emotional state a given producer of content wishes to impress on a message. In addition to this, social media networks have been arising as an ethereal space where countless people interact with each other and produce a hugeness of contents every minute. These contents, as cheap and easy to obtain, sustain a good base of research to study human behaviour, and in the case of this work, the human behaviour regarding emotions. Hereupon, this work aims at studying human emotions leveraging on social media contents.

Having this goal in mind, this dissertation proposes on building a method for inferring the emotional semantics, in terms of scores of valence, arousal, and dominance for textual contents (i.e., words and short texts). Later on, leveraging on the aforementioned method and social media contents from Twitter, are proposed two applications: the first one tries to assess well-being levels for populations across continental USA (excluding Alaska), and the second one aims at studying the relationship between the emotional content of a given message in Twitter and how this message spreads in the referred social network, e.g., trying to see whether markedly negative messages reach a larger number of users than neutral messages.

Keywords: Natural Language Processing, Sentiment Analysis, Information Diffusion, Social Media,

Resumo

As emoções estão presentes no dia-a-dia dos seres humanos e são consequentemente reflectidas na forma como os humanos comunicam e interagem entre si. Gestos, prosódia e as palavras que escolhemos num dado contexto são também bons reflectores do estado emocional que um dado produtor de conteúdos decide imprimir numa mensagem. Por outro lado, as redes sociais têm surgido como um espaço etéreo onde um número incontável de pessoas interagem e produzem uma imensidão de conteúdos a cada minuto. Estes conteúdos, sendo baratos e fáceis de obter, podem sustentar uma base de investigação relativa ao comportamento humano, e no caso deste trabalho, o comportamento humano no que concerne o uso de emoções. Posto isto, este trabalho ambiciona estudar as emoções humanas fazendo uso de conteúdos provenientes das redes sociais.

Tendo este objectivo em mente, esta dissertação propõe-se primeiramente a construir um método que infira a semântica emocional de conteúdos textuais (i.e., palavras e textos curtos), em termos das pontuações relativas a valência, excitação e dominância. Posteriormente, fazendo uso do método acima referido e em contéudos provenientes do Twitter, são propostas duas aplicações: a primeira tenta aferir níveis de bem-estar para populações dos estados continentais dos EUA (excluindo o Alasca), enquanto que a segunda aplicação se propõe a estudar a relação existente entre o conteúdo emocional de uma dada mensagem do Twitter com a forma como essa mesma mensagem se difunde na referida rede social.

Keywords: Processamento de Língua Natural, Análise de Sentimento, Difusão de Informação, Redes Sociais,

Agradecimentos

Gostaria de agradecer ao conjunto de pessoas que no seu todo contribuíram para este trabalho, directamente e indirectamente, com o seu apoio nas mais diversas vertentes. Destaco, sem qualquer ordem entre si, as pessoas que estiveram mais ligadas a esta etapa. Nomeadamente, o meu orientador, o Prof. Bruno Martins, devido à sua alta disponibilidade, compreensão, e competência na orientação do trabalho levado a cabo nesta tese. Destaco ainda o apoio, tolerância, companheirismo e investimento por parte da minha namorada e família, essencial para manter o foco, motivação e energia durante todo o percurso, assim como pelo seu voto de apoio nos momentos mais desmotivantes. Reservo ainda uma palavra de especial apreço aos amigos, que fora e dentro da faculdade, contribuíram para tornar este longos meses mais agradáveis. Por último, mas não menos importante, agradeço o apoio financeiro providenciado pela Fundação de Ciências e Tecnologias, na forma de bolsa de Mestrado inserido no projecto KD-LBSN, para a elaboração do trabalho de investigação descrito na presente dissertação.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Emotions play a major role in human interactions, and are, therefore, intrinsically embedded in either textual or oral communication between human beings. The emotional characteristics of a message, contribute consequently, to its semantics.

In literature, the computational analysis of the emotions expressed in messages is usually referred to as sentiment analysis. Sentiment analysis deals with the computational treatment of expressions of opinion, sentiment and emotion/affect, as expressed over natural language. The topic currently attracts significant attention from both academia and industry, e.g., due to the large amounts of user-generated contents that are nowadays available, from blogs, forums, micro-blogs and social media in general.

Social media, is nowadays, a topic of large interest mainly due to the fact that is an easy and cheap way of analyzing human behaviour and interaction. It is also a mean of capturing more natural human interactions when compared to inquiries, and also relevant because of the existing amount of data available. For instance, several studies have been studying the relation between contents in social media, and real measurable metrics, such as the happiness of populations (Loff *et al.*, 2015) or the weather (Li *et al.*, 2014). User-generated contents from social networks, are produced every day and around the globe on an unprecedented scale. Having such a large set of contents, we can leverage on these to ask a myriad of questions. Another topic of great interest, has been leveraging on these contents and studying how they diffuse from user to user, and around the world.

The focus of this dissertation was developing a method for extraction of emotional scores (in a continuous scale) from words (i.e., both in English, Spanish, Portuguese, Italian and German), and textual messages, in terms of Valence, Arousal and Dominance. Then relying on this method

and on a dataset of tweets, were produced two applications: (1) Inferring well-being of populations across USA, and (2) studying the relation between these scores, and the way the messages associated to the aforementioned scores spread in Twitter, according to some diffusion metrics (e.g., geographic coverage of a message).

1.1 Thesis Proposal and Validation Plan

In the context of my M.Sc. thesis, I propose to evaluate a method for extracting emotional scores (i.e., in a continuous scale of $[1 - 9]$) from words and texts, in terms of valence, arousal and dominance. More specifically, I resort to a regression model, that based on embeddings (i.e., vectors of dimension n, which capture the semantics) of words and texts, predicts for each one of the aforementioned emotional dimensions a score between $[1, 9]$. The aforementioned regression model is trained by associating for each word present in English emotional lexicons (i.e., dictionaries that associate words in English to scores of Valence, Arousal and Dominance, such as ANEW (Bradley & Lang, 1999) and the lexicon from Warriner *et al.* (2013a)), its corresponding word embedding to the corresponding emotional scores in the lexicons. This same method is also applied in a bilingual approach, i.e., regression models are trained leveraging on emotional lexicons in English such as ANEW (Bradley & Lang, 1999; Warriner *et al.*, 2013a), and are then used for predicting emotional scores for words in other languages (i.e., Spanish, Portuguese, Italian and German). Finally, was followed the same approach for predicting emotional scores for shorts texts (e.g., paragraphs, tweets, forum posts, etc.). When having embeddings associated to these same texts, and using the aforementioned regression models, if the embeddings associated to the texts are in the same space than the embeddings used for words when the models were trained, is possible to predict emotional scores, in terms of valence, arousal and dominance for the aforementioned texts.

So that the aforementioned methods can be evaluated, the mean absolute error (MAE) and the Pearson correlation coefficient ρ between the predictions and the scores in the lexicons, were measured. In the case of predicting emotional scores for words in English, the model was trained with words from Warriner *et al.* (2013a) and tested against words from ANEW (Bradley & Lang, 1999), and then trained with words from ANEW and tested against words from Warriner *et al.* (2013a). Similarly, in the case of predicting scores for other languages, the models were trained with words from Warriner *et al.* (2013a) and used to predict scores for words from lexicons in Portuguese, Spanish, Italian and German (Montefinese *et al.*, 2014; Redondo *et al.*, 2007; Schmidtke *et al.*, 2014; Soares *et al.*, 2012). Finally, in order to evaluate the method for estimating emotional scores, in terms of valence, arousal and dominance for short texts, models built using

words from Warriner *et al.* (2013a) were used to predict emotional scores as stated in datasets composed of short texts, such as the ANET (Bradley & Lang, 2007), the EmoTales (Francisco *et al.*, 2012), and one dataset composed of forum posts (Paltoglou *et al.*, 2013).

In addition to the aforementioned methods, which were the focus of the work developed during this dissertation, I also propose two different applications of the method for predicting emotional scores for texts. Both applications rely also in a dataset composed of tweets from the year of 2012. The main goal of these two applications is the study of emotions, through the usage of contents from social media.

The first application consists in predicting the well-being of populations across continental USA states (excluding Alaska), by taking into account all the embeddings of all the tweets issued from a given state, as well as, all the scores predicted to these same embeddings. As a ground-truth for this experiment is used the Gallup-Healthways composite well-being index relative to the year of 2012.

Regarding the evaluation of the first application, i.e., predicting the well-being of populations, a leave-one-out cross validation evaluation methodology was followed, and then the model was evaluated by measuring the MAE, the Root Mean Square Error (RMSE), the Pearson correlation coefficient ρ, and the Kendall correlation coefficient τ, between the predictions and the ground-truth scores (i.e., the Gallup-Healthways composite well-being index relative to the year of 2012).

The second application regards the study of the correlation between the emotion a given message is found to evoke on readers (i.e., in terms of Valence, Arousal, and Dominance, as predicted by the aforementioned regression models) with the way this same message spreads in Twitter (i.e., some metrics associated to the diffusion of messages, such as the geographic coverage). This last application intends to see if particular dimensional emotions evoked from messages can be linked to characteristics on the diffusion of these same messages.

1.2 Contributions

The research made in the context of this thesis has produced the following main contributions:

- State-of-the-art results when predicting continuous emotional scores, in terms of valence, arousal and dominance for unseen words in English, and good results when predicting scores for other languages;
- A set of extended emotional lexicons, in terms of valence, arousal and dominance for Portuguese, Spanish, Italian and German;

- A method capable of predicting emotional scores, in terms of valence, arousal and dominance for short texts based on neural word embeddings;
- Two applications leveraging on the aforementioned method for predicting emotional scores of short texts, in which were used tweets for: (1) predicting well-being of populations, and (2) studying the relations between the characteristics on the diffusion of a given message and the emotional reaction this same message evokes on readers.

1.3 Document Organization

The rest of the document is organized as follows: Chapter 2 describes some underlying fundamental concepts to this work, as well as, previous and related work. Chapter 3 presents a method that based on neural word embeddings predicts emotional characteristics of unseen words in English and in other languages (i.e., Spanish, Portuguese, Italian and German). Chapter 4, focus on the description of a methodology, that based on embeddings associated to short texts (e.g., sentences, paragraphs, etc) predicts emotional related ratings in a continuous scale that represent the emotions these same texts evoke on readers. Then, Chapter 5 leverages on the method presented in Chapter 4 to predict emotional ratings for tweets and: (1) tries to assess well-being of populations across continental USA (excluding Alaska). And finally, (2) studies the relationship between these emotional ratings and some metrics associated to the diffusion on Twitter, of the tweet that originated those predictions. Finally, Chapter 6 summarizes the main aspects discussed in this document, and points some possible future work directions.

Chapter 2

Fundamental Concepts and Related Work

This chapter describes, on Section 2.1 some underlying fundamental concepts, namely concepts associated to computation representation of text and those associated to complex networks and information propagation. Then, on Section 2.2, are presented some previous and related work, specifically those related with the extraction of emotions from social networks contents, and those related with the analysis on the diffusion of contents in social networks.

2.1 Fundamental Concepts

This section presents the underlying concepts for my work. Particular attention is given to ways of representing text for computational analysis, in Section 2.1.1, and to concepts related to complex networks and information propagation, in Section 2.1.2.

2.1.1 Representing Text Documents for Computational Analysis

During the development of this work, the need to represent text documents in a form that allows easy retrieval and/or comparison of the similarity among them, will arise. In order to address these needs, one can leverage the vector space model. This model represents each document

$\vec{d}$ as a vector of terms t, such that $\vec{d} = (t_1, t_2, ..., t_n)$. Each dimension of this vector corresponds to a given term from the vocabulary used throughout the document collection. In a vector $\vec{d}$ associated to a given document, the value of each dimension may be assigned in several ways. The value may, for instance, correspond to the number of times that the term appeared in the document, i.e., its frequency over the document. A common strategy is to assign these values by making use of the Term Frequency-Inverse Document Frequency scheme.

Term Frequency-Inverse Document Frequency (TF-IDF) indicates how important a term is in a document, considering also the entire document collection. The document for which the importance of the term is being evaluated is known to be contained in a larger collection of documents. TF-IDF grows proportionally to the amount of times that a given term appears in a document, and inversely proportional to the number of times that it appears in the corpus of documents. This way, when TF-IDF is applied to the English language, very common words such as *the*, *and* and *a* would not be considered as important.

The $\text{tf-idf}(t, d, D)$ function, expressed in Equation 2.1, receives as arguments a term t in a document d present in a corpus D.

$$\text{tf-idf}(t, d, D) = \text{tf}(t, d) \times \log\left(\frac{N}{d \in D : t \in d}\right) \tag{2.1}$$

In the formula, $\text{tf}(t, d)$ is the frequency of t in d, and N corresponds to the number of documents in the corpus D. The parameter $d \in D : t \in d$ corresponds to the number of documents $d \in D$ that contain the term t.

In order to compare the similarity of two documents, we may measure the angle between their corresponding vectors $\vec{d_1}$ and $\vec{d_2}$. The similarity between documents should increase as the angle between their associated vectors decreases. In practice, one can compute the cosine of the angle. This approach is also useful to retrieve the k documents which are more similar to a given set of terms. Following the described procedure, we should select the k documents with the least angle between their associated vector and the vector composed of the intended terms. Equation 2.2 shows how we can compute the cosine similarity between two vector representations $\vec{d_1}$ and $\vec{d_2}$.

$$\text{sim}(\vec{d_1}, \vec{d_2}) = \frac{\vec{d_1}.\vec{d_2}}{\vec{d_1} \times \vec{d_2}} = \frac{\sum_{i=1}^{n} \vec{d_1}(i) \times \vec{d_2}(i)}{\sqrt{\sum_{i=1}^{n} \vec{d_1}(i)^2} \times \sqrt{\sum_{i=1}^{n} \vec{d_2}(i)^2}} \tag{2.2}$$

2.1.2 Complex Networks and Information Propagation

The two main constituents of a social network are the users and the relations among them. We can model, in a graph G, users as nodes V and the relations among them as edges E. More formally, a graph $G = (V, E)$ is composed by a set nodes V and a set of edges E. Every edge $e_{ij} \in E$ connects two nodes $v_i, v_j \in V$. An edge may optionally have an associated direction, being a directed edge. Graphs with directed edges are called directed graphs. In contrast, those with undirected edges are called undirected graphs. A weight can also be associated to a given edge, and this may be used to represent a given property considering the connection between two nodes. Regarding the computational representation of graphs, it can be made through adjacency lists or through a neighbourhood matrix. Adjacency lists are more suitable for sparse graphs (i.e., graphs where the number of edges is much less than the possible maximum) while adjacency matrices are best for representing dense graphs (i.e., graphs where the number of edges is very close to the possible maximum).

Graphs are typically analysed through properties computed over their nodes and/or edges. One of them, the degree k_v of a node v, denotes the number of edges $e \in E$ connected to the node v. If the graph is directed, the degree may be measured in relation to the direction of incidence. A parameter k_v^{in} can be used to refer to the number of incident edges on node v, and k_v^{out} to the number of edges pointing out of v. The degree is useful to identify the most important nodes in a social network. The most important users are usually the most connected ones, so the degrees associated to their nodes are usually the highest ones.

The degree distribution in a graph, $\mathrm{P}(k) = \frac{n_k}{|V|}$, where n_k is the number of nodes with degree k and where V is the set of nodes, is another interesting property to analyse. Network models based on random graphs can be shown to exhibit a binomial distribution in terms of the node degree, while graphs corresponding to social networks typically follow a power law distribution. In social network graphs, most of the nodes have a small degree, and the number of nodes with a given degree decreases as a power of the degree.

In order to assess how central a node is in a network, one can also use a property commonly referred to as betweenness centrality. This indicator is linked to the influence a given node has in a network. A node whose betweenness centrality is high has an important role in the forwarding of messages that move around the network. The formula that allows us to calculate the betweeness

centrality of a node v is given in Equation 2.3.

$$\mathrm{B}(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2.3}$$

In Equation 2.3, $\sigma_{st}(v)$ corresponds to the number of shortest paths between nodes s and t that pass through v, and σ_{st} corresponds to the total number of shortest paths between nodes s and t.

In a graph, the nodes may also be more or less clustered together. To measure the extent to which the nodes are clustered/tied together, we may calculate the global clustering coefficient. Equation 2.4 formalizes this coefficient.

$$C = \frac{1}{|V|} \sum_{i=1}^{|V|} \frac{\text{number of pairs of neighbours of } i \text{ that are connected}}{\text{number of pairs of neighbours of } i} \tag{2.4}$$

The clustering coefficient provides an understanding of the extent to which, in a given graph, we have communities of nodes that are highly connected among themselves. Romero *et al.* (2011) studied the relationship between the factor of clustering existent in a set of early adopters of a given highly diffused Twitter hashtag (i.e., first m individuals to mention a given hashtag) and the category (e.g., political, sports, etc) in which this same hashtag lies within. Romero *et al.* (2011) found that the subgraphs composed of early adopters of political hashtags, SG_p, tend to have a greater amount of strong ties between nodes than the subgraphs composed of early adopters of other types of hashtags, SG_o. Nevertheless, the clustering coefficient of SG_p was lower than the overall clustering coefficient for all subgraphs composed of early adopters of other hashtags. The authors concluded that this high aumount of strong ties in SG_p comes from a core of strong ties inside a larger core of weak ties.

Information propagation processes, in the context of a social network like Twitter, model how messages originated by a given source propagate from user node to user node. One such process may be defined as *the set of influence paths that share a common root*. In this definition, the root is the user who first posted the message (or tweet, in the context of Twitter) which is going to be diffused (or retweeted, in the context of Twitter) by other users in the network. The set of influence paths is an ordered set of directional edges between two or more user nodes, that express the chain of influence originated when the root initially posted a message and finished when the last user in the process diffused that same message. By influence, it is meant the relationship between two users, in which the *influencer* is the user that exposes a message to the *influenced user* and influenced the *influenced user* to propagate the message. In Twitter, the *influencer* is the user from whom the *influenced user* retweeted a given message.

When the exact influence path of the information propagation processes is not totally explicit, due to limitations of the data sources (i.e., in Twitter the user from whom the message was retweeted is not explicit) or due to missing data (i.e., missing retweets in a given sample of the Twitter data stream), the assignment of the influence relationships between users must be done in function of an influence assignment model. More concretely, since in a given online social network we have that a user can be exposed to a message and influenced to share it by a broad set of users (e.g., in the context of Twitter, the users he follows or any other user that publicly shares his content), one of these users must be assigned as the *influencer*. To suit these needs, Taxidou & Fischer (2014) proposed the following influence models: (1) users are influenced by the first exposure of the message, (2) users are influenced by the last exposure of the message, (3) users are influenced by the most followed user, or (4) the users are influenced by the user whose messages are forwarded more times. Each one of these models, by itself, is not totally accurate with respect to the real diffusion process they intend to model, since these models are simplistic abstractions.

Another issue that arises when reconstructing information propagation processes, from datasets of messages collected from social networks like Twitter, is that one may have missing messages (i.e., retweets). These missing messages imply that nodes and influence paths will be equally missing from the graph that represents the reconstructed diffusion process. Therefore, it may happen that the model of influence employed to reconstruct the diffusion process cannot assign a previous influencer to a given retweet, and thus, that given retweet may become the root of a new fragment of the graph that represents the reconstructed diffusion process. Consequently, one may end up with a representation of reconstructed diffusion process composed by multiple disconnected graph fragments. Still regarding this issue, Sadikov *et al.* (2011) studied the effect of missing data on a diffusion process, and they proposed a numerical method that attempts to correct it. This method, given a k-tree model of an incomplete diffusion process, returns properties of the complete process which this incomplete process refers to. The authors showed experimentally that the referred method can correctly estimate properties of a diffusion process when having 90% of missing data.

Consider $C(U, E)$ to be a graph of influence paths between users, that represents a reconstructed information propagation process of a given tweet, with U representing all the users nodes who retweeted the message, and with E denoting the edges that represent the influence relationships between users which were assigned with regard to a specific influence model. In order to assess the connectivity of an information propagation process reconstructed through a given influence model, Taxidou & Fischer (2014) presented two metrics: the connectivity rate and the root fragment rate. The connectivity rate (CR) gives the percentage of users that are influenced

by another user (i.e., the percentage of nodes that are connected to at least one other node in C), and is calculated as denoted in Equation 2.5.

$$\mathrm{CR} = \frac{|\{u|(u', u) \in E \vee (u, u') \in E\}|}{|U|} \tag{2.5}$$

The metric presented in the previous equation only provides insights into how connected are the nodes in C, but one can imagine a scenario where a large percentage of the nodes U are connected and there are not paths that connect the root node to the nodes in the bottom of the diffusion process. Thus, Taxidou & Fischer (2014) also presented the aforementioned root fragment rate (RFR) metric. The root fragment rate metric gives the percentage of nodes that are directly connected to the root user, or connected to the root through an influence path composed by one or more users. This metric may be calculated as expressed in Equation 2.6, where u_r, denotes the root user.

$$\mathrm{RFR} = \frac{|\{u_j \in U|\text{iff exists a path } u_r, ..., u_j \text{ in C}\}|}{|U|} \tag{2.6}$$

In the context of the diffusion of messages in networks, Romero *et al.* (2011) also presented two metrics (stickiness and persistence) to characterize the information propagation process originated by a given piece of information. The stickiness metric may be perceived intuitively as the probability that a given piece of information will be diffused by a user when he is exposed to it, while the persistence metric may be intuitively conceived as the rate of decay of the diffusion of a piece of information after its diffusion peak. Romero *et al.* (2011) based the definition of these two metrics on a function $\mathrm{D}(k) : [0, K] \rightarrow [0, 1]$, which returns for each k, the fraction of users in a given network who spread the considered piece of information right after their k^{th} exposure to it, when considering K maximum possible exposures. The stickiness S of a given piece of information can be formally defined as the maximum value of the function $\mathrm{D}(k)$, as stated in Equation 2.7.

$$S = \max_{k \in [0,K]} \mathrm{D}(k) \tag{2.7}$$

On the other hand, the persistence P may be defined as expressed in Equation 2.8, where A represents the area under the curve (see Fig. 2.1) formed while plotting the function $\mathrm{D}(k)$ by connecting with a straight line the point $\mathrm{D}(k)$ to the point $\mathrm{D}(k + 1)$ In the formula, the parameter K refers to the considered number of maximum possible exposures.

$$P = \frac{A}{K \times \max_{k \in [0,K]} \mathrm{D}(k)} \tag{2.8}$$

Ghosh & Lerman (2011) formally characterize a cascade by the use of a cascade generation

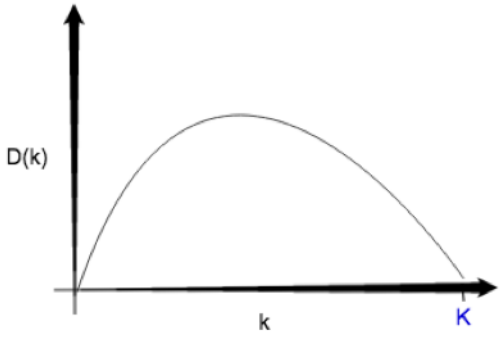

Figure 2.1: Graphical representation of the concept of persistence

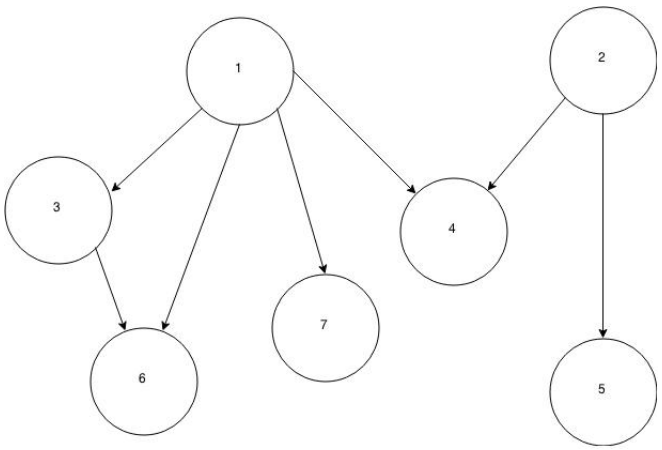

Figure 2.2: Representation of a graph containing two diffusion processes, adapted from Ghosh & Lerman (2011)

function $\mathrm{c}(j, \alpha_{j,i})$, which encodes a given diffusion process in a graph $G(V, E)$. This function is parametrized by a transmission rate $\alpha_{j,i} \forall j, i \in [1, |V|]$ which represents the probability of a node i influenced at a time t_i influencing a node j at time t_j, having $t_j > t_i$. The authors assume, for simplicity, that $\alpha_{j,i}$ is equal for any given value i or j and that the nodes V in the graph G are labeled according to their order of *infection* regarding the diffusion process. Thus, $\mathrm{c}(j, \alpha_{j,i})$ represents the cascade at time t_j. From now on, I will use the cascades represented in Figure 2.2 to support the following explanations. Notice that the example from Figure 2.2 represents a diffusion process with two different roots and in which some nodes may be influenced by more than one node (e.g., node 6 is influenced by the nodes 1 and 3). This function definition requires the initial value of the cascade to be set as a constant, i.e., at the time nodes 1 and 2 (Fig. 2.2) are active, and if we consider that these nodes are roots of cascades, the values of the function $\mathrm{c}(1, \alpha)$ and $\mathrm{c}(2, \alpha)$ are set as a constant. I assume, for explanation purposes, $\mathrm{c}(1, \alpha) = \mathrm{c}(2, \alpha) = 1$. The

cascade function value at the time node j is infected (t_j) is recursively calculated as expressed in Equation 2.9, where K denotes the number of cascades in the diffusion process.

$$\mathrm{c}(j,\alpha) = \sum_{i \in \mathrm{neighbours}(j)} \alpha \mathrm{c}(i,\alpha) = \sum_{p=1}^{K} \mathrm{f}(j, i_p, \alpha)\mathrm{c}(i_p, \alpha) \tag{2.9}$$

Using the previous equation, and considering the example from Figure 2.2, we have that $\mathrm{c}(4,\alpha) = \alpha\mathrm{c}(1,\alpha) + \alpha\mathrm{c}(2,\alpha)$ and $\mathrm{c}(6,\alpha) = \alpha\mathrm{c}(1,\alpha) + \alpha\mathrm{c}(3,\alpha)$. Since $\mathrm{c}(3,\alpha) = \alpha\mathrm{c}(1,\alpha)$, $\mathrm{c}(6,\alpha) = \alpha\mathrm{c}(1,\alpha) + \alpha^2\mathrm{c}(1,\alpha)$. In the previous equation $\mathrm{f}(j, i_p, \alpha)$ encapsulates the cumulative effect of the cascade seeded at i_p over node j, having $t_j > t_{i_p}$ (e.g., $\mathrm{f}(6,1,\alpha) = \alpha + \alpha^2$ and $\mathrm{f}(6,2,\alpha) = 0$).

Ghosh & Lerman (2011) leverage the computation of cascade properties on the cascade generation function. For example, having $\mathrm{c}(i_p,\alpha) = 1$, $\mathrm{f}(j, i_p, 1)$ gives the total number of paths from the root i_p to j (e.g, $\mathrm{f}(6,1,1) = 2$). The cascade generation function can also be used to compute the total length $\mathrm{l}(j, i_p)$ of the paths from the root i_p to j. Equation 2.10, using Leibniz's notation, shows that one should compute $\mathrm{l}(j, i_p)$ as the derivative for function $\mathrm{f}(j, i_p, \alpha)$ at the point where $\alpha = 1$.

$$\mathrm{l}(j, i_p) = \left.\frac{d\mathrm{f}(j, i_p, \alpha)}{d\alpha}\right|_{\alpha=1} \tag{2.10}$$

For example, the total length of the paths from the root node 1 to the node 6, is $\mathrm{l}(6,1) = 3$, since from node 1 to 6 there exist two possible paths, namely $1 \to 3 \to 6$ and $1 \to 6$ whose total total length is 3. The same line of thought can base the computation of the average path length l_{av}, of K cascades in G, which can be computed as stated in Equation 2.11.

$$l_{av} = \frac{\sum_{j \in V} \sum_{p=1}^{K} \mathrm{l}(j, i_p)}{\sum_{j \in V} \sum_{p=1}^{K} \mathrm{f}(j, i_p, 1)} \tag{2.11}$$

Finally, the length of the longest path of any cascade $l_{\max}$ in G, commonly referred as the diameter of a diffusion process, may be calculated as expressed in Equation 2.12.

$$l_{\max} = \max_{j \in [2,N]} \frac{d \min_{i \in \mathrm{neighbours}(j)} \alpha \mathrm{c}_{\min}(i,\alpha)}{d\alpha} \frac{\alpha}{\min_{i \in \mathrm{neighbours}(j)} \alpha \mathrm{c}_{\min}(i,\alpha)} \tag{2.12}$$

In subsequent work, Lerman *et al.* (2012) presented informally a set of properties which characterize diffusion processes, such as: (1) the size, which corresponds to the number of nodes which are influenced by a given diffusion process, (2) the maximum (formalized in Eq. 2.12) and minimum diameter, which can be translated respectively to the length of the longest chain of influence in the given diffusion process and to the length of longest shortest path from the root to all nodes in the given diffusion process, and (4) the spread, which is the maximal branching

factor of influence in the given diffusion process, i.e., the maximum number of nodes which any node in the diffusion process influences.

2.2 Previous and Related Work

This section summarizes recent work related to different aspects of my thesis namely, in Section 2.2.1, works related with the extraction of emotions from contents present in social media networks (i.e., the emotions these contents are found to evoke on users when exposed to them). In Section 2.2.2 are presented some works related with the study of how social media contents spread in social networks, particularly focusing on studies that attempted to analyze spatio-temporal properties of information propagation.

2.2.1 Extracting Emotions from Social Media Contents

This section presents previous work concerning with the extraction of emotion and opinion-related information from social media contents. Particular focus is given to the methods used to extract the emotions evoked on users when exposed to the contents. There are also presented some works that try to relate the emotions extracted from social media contents with, for instance the weather (Li *et al.*, 2014), or the well-being of populations (Loff *et al.*, 2015).

Temporal Patterns of Happiness in a Global Social Network

Dodds *et al.* (2011a) studied textual expressions posted in the context of Twitter messages in order to discover and analyze their temporal variation in terms of happiness. To measure happiness, the authors built a *tunable, real-time, remote-sensing, and non-invasive, text-based hedonometer*[1].

The authors ran a survey using Amazon's Mechanical Turk to collect emotional evaluations for approximately ten thousand individual words. Each individual word was subjected to fifty individual evaluations, and each evaluator was requested to assign a score between 1 and 9 for the emotional valence of the words, with 1 corresponding to negatively charged words, and 9 corresponding to positive words. This interval allowed the authors to remain consistent with the

[1] *http://hedonometer.org/*

evaluation from the well known study named Affective Norms for English Words (ANEW), by Bradley *et al.* (1999). The creation of the list of words to be evaluated was conducted in the following way:

1. Words were collected from: (1) a large Twitter dataset, (2) Google Books in English, (3) music lyrics dated from 1960 to 2007, and (4) articles in The New York Times, published between 1987 and 2007.

2. For each one of these sources, words were ordered by their occurrence frequency. The top 5000 most frequent words of each source were merged into a single set containing a total of 10.222 words.

A reassuring indicator for the robustness of this evaluation of words is the result of the correlation with the valence values from Bradley *et al.* (1999). The authors measured a Spearman correlation of 0.9444 and a $p_{value} < 10^{-10}$.

The main goal behind the work of Dodds *et al.* (2011a) was to be able to sense happiness levels at a societal level and in near real-time, solely making use of the analysis of data collected from Twitter. This is based in the authors belief that Twitter's users frequently express their current state of happiness instead of a contemplative evaluation of their overall life.

The dataset of tweets used to sustain an initial study was collected between September 9, 2008 and September 18, 2011. This dataset included 46.076 billion words in 4.586 billion tweets posted by around 63 million users.

The algorithm used to assess the happiness of tweets is based on: (1) the human evaluation of the happiness level of individual words, and (2) a simple method to extrapolate from these happiness levels for the words to the entire text. This method starts by calculating the frequency of the words in a text T. Then, the weighted average of happiness for a text T is calculated as follows:

$$\mathrm{h}_{\mathrm{avg}}(T) = \frac{\sum_{i=1}^{N} \mathrm{h}_{\mathrm{avg}}(w_i) f_i}{\sum_{i=1}^{N} f_i} \tag{2.13}$$

In the formula, f_i is the frequency of the word w_i, and $\mathrm{h}_{\mathrm{avg}}(w_i)$ is the estimate of happiness for word w_i. Words are only considered in this analysis if there exists a happiness estimate for them.

The hedonometer can be enhanced and tuned, by ignoring subsets of the list with 10.222 words. The subjacent idea is to remove from the list those words whose score is neutral. Doing so, the hedonometer will become more sensitive. For a given x, the authors tried to remove those words whose $\mathrm{h}_{\mathrm{avg}}$ obeys to the relation $5 - x < h_{avg} < 5 + x$. Through experiments, the authors found

that $x = 1$ was an adequate trade-off between sensitivity and text coverage.

The authors also noticed that the described methodology has some clear limitations. The first one is that the method is fragile for tiny texts, such as an isolated tweet. However, for the purpose of the authors' study, this is not an issue, since the sentiment is always inferred from a large set of tweets. The other limitation is that the proposed method does not extract meaning or context from the text, thus ignoring a considerable amount of useful information, and occasionally producing errors in the analysis. Once again, the authors consider that this limitation is overcame in the context of big datasets.

Some key findings, worth of mention, when relating the information extracted with this hedonometer with temporal properties are as follows:

- The day of Bin Laden's death was considered by the hedonometer as a very sad day. This happened because the words associated to this event were markedly negative. This suggests that every measurement using the hedonometer must be accompanied by the analysis of the most prevalent words of the analyzed set of data.
- Saturday is, in average, the happiest day of the week, followed by Friday and Sunday. Tuesday is the saddest day in the week. This reading goes against the common-sense idea that Mondays are the worse days.
- The happiest hour of the day is from 5am to 6am. There is an accentuated drop in happiness until 12am, followed from a gradual decrease until the lowest hour of the day (10 to 11pm). After that, there is a fast recovery, during the dawn, until happiness reaches its maximum point.

Using the hedonometer it is also possible to measure happiness averages for tweets containing particular keywords. This may be useful for applications such as measuring the sentiment associated to a given company, to a given product or service, or to a given individual or organization.

In a subsequent study, Bliss *et al.* (2012) made use of the hedonometer described by Dodds *et al.* (2011a), in order to study the relationship between users' happiness and their connections in the Twitter social network. The main question behind this study was to see *whether happiness is assortative in reciprocal-reply networks*. The authors of this study also:

- Tested the hypothesis proposed by Christakis & Fowler (2013) that states that *assortativity of happiness may be detected up to three links away*;
- Sought to understand how happiness spreads over social networks;

- Attempted to understand how social environment properties influence user happiness and vice-versa.

The authors discarded the usage of a simple directed network, based on replies between tweets, to model the analysed problem. They did so because replies do not imply reciprocity between users. If a user i replies to user j, there is absolutely no evidence that j has read or taken into account this reply. The authors state the following, as a *sine qua non* condition, for two users being connected in a reciprocal-reply network: user i must have replied to j, and user j must have replied to i at least a single time in a given period of time. The reciprocal-reply network is represented by the authors in the form of a graph $G(V, E)$, where the nodes V correspond to users, and where the edges E correspond to reciprocal-only connections between users.

Two issues emerged when the authors tried to model this network: (1) how can one know until when is it correct to maintain a connection between two users, since this situation may lead to inaccurate readings, and (2) how can one deal with the huge accumulation of information during the time, when one intends to measure assortativity between user nodes in relation to a measure inferred from the shared messages (e.g., happiness). The authors tried to solve these problems by analyzing the networks at small temporal scales and calculating the users' happiness based only on the tweets posted in that temporal scale.

To sustain this study, the authors collected around 100 million tweets from the Twitter streaming API service, during the period between September 2008 and February 2009. They estimate to have collected around 38% of all tweets produced in that period. The relevant information to be extracted from each tweet is the id of the message, the id of the author, the id of the message being replied to, and finally the id of the user being replied to. Since Twitter did not allow the collection of all tweets published in the mentioned period, there are some flaws in this dataset. As a result of this, there may exist some user nodes that are not connected when in reality they should, or they may be connected through a larger path than the real one. It is also relevant to mention that users may interact with each other without making use of Twitter's reply function.

In order to measure users' happiness, the authors resort to the use of the hedonometer. After calculating each user's happiness score, they create pairs (h_{v_i}, h_{v_j}), where h_{v_i} and h_{v_j} denote the happiness scores of nodes v_i and v_j respectively. Spearman's correlation for these pairs is then calculated, and this correlation represents the similarity of happiness between user nodes that are neighbours in the network. Then, the authors also attempted to measure how strong is this correlation between users that are two and three links away.

The authors found that: (1) the happiness correlation between users nodes decreases when the path length between them increases, thus concluding that the *network is assortive with respect*

to happiness, and (2) the average happiness increases in function of the degree of user nodes. Degree is defined as the number of incident edges on the user's node.

To better examine the foundation of these observations, the authors tested them against a null model. This null model preserves the network topology and randomly permutes the happiness scores associated to each node. When the authors applied Spearman's correlation to this null model, they found no relationship between neighborhood and happiness. This result reinforces the validity of the aforementioned results. The authors also tested if the discovered correlations were caused by the similarities in the words employed by users that were neighbors, but no evidence was found for this fact.

Exploring Mood-Weather Relationships from Twitter

Li *et al.* (2014) studied the correlation between different facets of human mood with meteorological effects, leveraging on messages collected from Twitter. The authors relied on the assumption that user-generated content in online social networks like Twitter reflects, directly or indirectly, the mood of the users that posted that content. The authors also rely on the belief that the data errors that may exist in the source of the information supporting the study, are attenuated by the large amount of data that is processed.

The authors supported their study on two distinct data sources:

- A dataset belonging to Carnegie Mellon University, containing geo-tagged tweets gathered from 32 urban areas of the USA, corresponding to around 2% of all the tweets posted on Twitter in 2010 and 2011;
- Meteorological observation data from the National Oceanic and Atmospheric Administration (NOAA), which contains meteorological factors like average temperature, daily precipitation, total solar energy received, etc.

The mood of individuals is naturally also affected by other events besides the weather, such as political events, cultural events, etc. In order to identify and filter-out the messages from Twitter that are associated to these kinds of events, the authors propose an approach based on machine learning. The authors started by employing the system by Ritter *et al.* (2012) in order to extract open domain public events from Twitter. The aforementioned system uses an approach based on Condition Random Fields, as described by Lafferty *et al.* (2001), to identify named entities and event phrases. For event clustering, all these named entities and event phrases, which were identified by the aforementioned approach, were then introduced into a Latent Dirichlet Allocation

model (Blei *et al.*, 2003). The clusters mined from the Latent Dirichlet Allocation model are then manually identified and labeled. The number of clusters, which were found to correspond to related event types, was 52. Lastly, the tweets containing event-related mentions were filtered.

For mining opinions, the authors used the lexicon from OpinionsFinder[1], which contains positive and negative terms. Since OpinionFinder only distinguishes between positive and negative terms, the authors also used the terms from the Profile of Mood States (POMS) questionnaire by Pollock *et al.* (1979). This questionnaire uses a scale to measure transient emotional states. The authors extended the original term list from the POMS questionnaire, by identifying the words that co-occurred most with each given term. To accomplish this task, they resorted to a collection of 4 and 5-grams collected from Google and previously described by Bergsma *et al.* (2009).

Tweets not containing terms from OpinionFinder or from the POMS extended term list are discarded. It is important to mention that not every tweet containing mood indicating terms expresses the mood of the author (e.g., *I feel like having McDonald's for lunch*). Taking the aforementioned problem into account, the authors trained a maximum entropy classifier to distinguish between positive, negative and neutral tweets. This classifier takes into account several features (e.g., sentiment indicators, POS of the sentiment indicators, context words of the sentiment indicators and correspondent POS, etc.). By sentiment indicators the authors mean the tokens of the tweet that correspond to a mood indicating term. The POS of each term was extracted using the system previously described by Owoputi *et al.* (2013).

To assess the sentiment score x_t of a given the day t, the authors calculate the ratio of positive versus overall messages, as described in Equation 2.14.

$$x_t = \frac{\mathrm{count_t}(positive)}{\mathrm{count_t}(positive) + \mathrm{count_t}(negative)} \tag{2.14}$$

In order to capture the non-linear relationship between mood and weather, the authors applied a Generalized Mixed Model as described by Hastie & Tibshirani (1990) and by Wood (2011). This model takes into account factors such as (1) time auto-correlation, (2) the inter-correlation between the regression variables, and (3) external information added by additional variables.

The authors presented two sets of results. The first set focused on seeing if a given meteorological factor has a positive or negative contribution to the mood of the population in study (e.g, temperature does not make a significant contribution to mood state, more snow leads to a negative mood state, etc). The other set of results also explores the influence of weather in the mood, but explores how weather factors contribute to the variation of mood in relation to the transitional

[1] http://mpqa.cs.pitt.edu/opinionfinder/

emotion states used in POMS (e.g., the hotter the angrier, high temperature leads to tiredness, cool temperature leads to slepiness, etc).

Characterizing Geographic Variation in Well-Being Using Tweets

In the context of the World Well-Being Project[1], Schwartz *et al.* (2013b) approached the problem of predicting the geospatial variation of happiness patterns, understanding the factors that cause this variation and finding the language properties that characterize life satisfaction, by using tweets. The main approach was to map tweets to the US counties they were posted from, and then to correlate properties extracted from these tweets with life satisfaction indexes associated to these counties, as measured by life satisfaction phone surveys (Lawless & Lucas (2011)). To sustain this study, the authors collected around a billion tweets during the period between November of 2008 and January of 2010, via the Twitter *garden hose* API.

With the purpose of finding language patterns characterizing subjective well-being, the authors also report on the following main contributions:

- They hand-built a lexicon combining multiple word lists, including the words from the LIWC study, by Pennebaker *et al.* (2001), as well as the terms from the PERMA lexicon, by Seligman (2011). Each one of the word lists, forming this combined lexicon, is associated to a semantic and syntactic category (e.g., positive emotion, leisure, swear, pronouns, verbs, etc);
- The authors derived clusters of lexico-semantically related terms, automatically from the application of a Latent Dirichlet Allocation topic model (Blei *et al.*, 2003). The topics used in this study were the ones derived from Facebook status updates, as described by Schwartz *et al.* (2013a).

The per-county usage of the aforementioned lexical categories is measured by calculating the percentage of a county's words which are in a given category. To measure the probability of each topic within each county, the authors presented the following equation:

$$\mathrm{P}(topic|county) = \sum_{word \in topic} \mathrm{P}(topic|word).\mathrm{P}(word|county) \tag{2.15}$$

In the previous equation, $\mathrm{P}(word|county)$ denotes the normalized probability of that word occurring in that county, and $\mathrm{P}(topic|word)$ corresponds to the probability of a topic given a word. This last probability is granted by the Latent Dirichlet Allocation topic model.

[1] http://wwbp.org/

Due to size of the Twitter dataset, the authors used MapReduce to aggregate the words per county. Since only a small percentage of the tweets have an associated set of geographical coordinates, the authors developed a method to map the tweets which do not contain this information to a given location, by parsing the location text field associated to the user who posted the tweet. The developed method is based on a cascaded set of rules which map the string present in the location text field to a US County. Firstly, the method tokenizes the string and tries to match these tokens to country names. If a country that is not USA is matched, the tweet is discarded. Then, by using the token preceding the country (if existent), the method tries to match the city and state name. If only the name of the city can be matched, the method associates the city to the most probable state, according to the population size of all cities with that name. However, if the city name lies in the set of the top 100 largest non-USA cities, the tweet is discarded. All the others tweets are discarded.

Then, the authors extracted the usage of every given lexical category and topic, per county. With these features extracted, the authors ran a correlation analysis between all topics, all categories, and the life satisfaction scores, per county. In this correlation analysis the authors used the least squares linear regression over standardized variables method. This method produces a Pearson's r as a measure of the linear correlation between each two of the aforementioned variables. The results are only considered if they pass a Bonferroni-corrected p-value of 0.05 (i.e., p-value $< 0.05/2000$, since the authors considered 2000 topics). Finally, due to the high number of obtained correlation results between the topics and the life satisfaction scores, the authors used word clouds in order to visualize the topics and select those which seem to be more significant.

The predictive model for the variation of happiness within a given county uses, as features, the lexical categories and the topics extracted from the messages posted from that given county. The authors used as control variables the best know predictor of county well-being (e.g., median age, median household income, percentage of females, and percentage of minorities), as obtained from census data. They used these variables to verify whether their models, based on language, could improve a prediction model solely based on the control variables.

The authors specifically used a LASSO (least absolute shrinkage and selection operator) regularized linear regression model (Tibshirani, 1996), in order to promote a sparse usage of the considered features. The LASSO regularization approach pushes the features which are less predictive to be weighted as zero, deselecting them from the regression model. This regularization method was particularly useful in the context of this work since the size of the sample was smaller than the number of features.

The authors found that the aforementioned control variables can be more predictive than the topics. The topics are, in turn, more predictive alone than the lexical categories by themselves. The most accurate results were obtained when combining all these features (topics, lexical categories and control variables). These results confirm that the words in tweets can improve the accuracy of models for predicting state-level happiness which are solely based on the aforementioned control variables.

In a recent study taking inspiration from the work of Schwartz *et al.* (2013b), Loff *et al.* (2015) proposed a method to estimate population well-being over the USA. This method leverages on geo-referenced messages from Twitter (e.g., those which were published from mobile devices with GPS sensors) collected within the year of 2012, simultaneously with human evaluation of the emotions encoded in particular words, in the form of lexicons such as ANEW (Bradley *et al.*, 1999) and LabMT[1]. The authors learned a linear regression model using features corresponding to word counts in the aforementioned lexicons of emotionally charged terms, that estimates the Gallup-Healthways[2] composite well-being index based on telephone interviews with 1.000 people over USA, for the year of 2012.

Leveraging on the aforementioned lexicons, the authors computed for each Twitter message, the average score of the following dimensions: valence, arousal, dominance and overall happiness. The valence, arousal, and dominance averages leveraged on the ANEW lexicon, while the happiness average score was calculated with the LabMT lexicon. The authors compute the overall score of a given dimension dim, in a tweet, as expressed in Equation 2.16.

$$\dim(tweet) = \frac{\sum_{i=1}^{n} \dim(i) \times f_i}{\sum_{i=1}^{n} f_i} \tag{2.16}$$

In the aforementioned equation, f_i denotes the number of times that the ith word of the considered lexicon appears in the tweet's text, and the function $\dim(i)$ returns the average value of the considered dimension for the ith word within the considered lexicon.

Then, by aggregating the scores of each Twitter message according to geospatial regions (i.e., USA states excluding Alaska and Hawaii, or the corresponding counties), the authors extracted, for each one of these regions, a broad set of features, ranging from simple statistics (e.g., minimum, maximum, mean, mode, standard deviation, etc.) over the scores for each considered dimension, to more complex features, such as these same simple statistics but applied to scores whose calculation did not took into account words whose value regarding a given dimension

[1] http://neuro.imm.dtu.dk/wiki/LabMT
[2] http://www.healthways.com/

matched some criteria (e.g., average happiness score lying within with a delta of 1 from the neutral score of 5, or within a given interval). An aggregate of 46 features was hence extracted for each one of the aforementioned regions, and then used to train a linear regression model which tries to approximate the Gallup-Healthways state-level well-being index.

The linear regression model was learned with the Elastic Net approach for regularizing linear least squares regression models, proposed by Zou & Hastie (2005). The Elastic Net regularization approach overcomes the limitations of the LASSO approach, i.e., when having several highly correlated variables, the LASSO approach is likely to select one variable and ignore the others. Moreover, the Elastic Net approach is more useful compared to the LASSO approach if the number of considered features is larger than the number of training instances. Considering a linear regression model of the form $\vec{y} = X\vec{b} + \vec{e}$, where the vector $\vec{y}$ is the prediction outcome, the matrix X encodes all the considered features associated to each one of the training instances, the vector $\vec{b}$ being the regression coefficients, and finally the vector $\vec{e}$ representing the difference between the prediction outcomes and the observed response in the training data, the Elastic Net approach regularizes the model by solving the following optimization problem: $\vec{b} = \text{argmin}_{\vec{b}}||\vec{y} - X\vec{b}||^2 + \lambda_1||\vec{b}||_1 + \lambda_2||\vec{b}||_2^2$. In the aforementioned optimization problem, λ_2 and λ_2 weight the l_1 and l_2 regularization penalties, respectively. In order to find the model parameters, the authors resorted to an implementation from the glmnet[1] package for the R system for statistical computing, described by Friedman *et al.* (2010). For evaluating the quality of the constructed model, regarding the approximation to the Gallup-Healtways well-being index, the authors used a leave-one-out cross validation scheme. The authors report a Mean Absolute Error (MEA) of 0.92 and a Root Mean Square Error (RMSE) of 1.22. A baseline approach, consisting in assigning the average value for the Gallup-Healthways well-being index to all states, results in a MEA of 1.40 and a RMSE of 1.73. The authors also reported Pearson and Kendall correlation coefficients of 0.7441 and 0.5862, respectively. It is also worth to mention that the authors found that the most discriminative feature in the model was the happiness score whose calculation did not take into account the neutral words from the LabMT lexicon.

Language Usage and Influence on Twitter

Quercia *et al.* (2011) crawled 31.5 million tweets, analyzed them regarding the use of language by making use of the LIWC lexicon, and then studied the relationship between the language that the authors of these tweets employ and the types of users that post the messages (e.g., influential, popular, star, etc.). The authors attempted to verify the hypothesis that states that, in

[1] http://cran.r-project.org/web/packages/glmnet/index.html

Twitter, influential users mostly tend to express negative sentiments in their tweets. The authors also argue that the followed approach will also help to test whether users in Twitter are look-alike nodes in a graph, or whether they show significant linguistic differences between themselves.

Concerning the Twitter dataset that this study leverages on, and in order to reduce the variability in the use of language across countries, the authors decided to collect tweets issued only from the United Kingdom. Thus, the authors captured these tweets resorting to the Twitter Streaming API, and collected tweets during the period between the 27th of September 2010 and the 10th of December 2010, from 250.000 users who were followers of some British news outlet profile (e.g., followers of The Independent, The Sun, etc).

In order to characterize each user, concerning the use of language in the tweets he authored, the authors resorted to the LIWC lexicon. In the LIWC lexicon, words can fall into 72 abstract categories, and each word can belong to multiple categories. Previous authors, such as Gill *et al.* (2009), have found that 10 of these categories (e.g., *first person, second person, cognitive, positive/negative emotion, etc.*) correlate with personality characteristics. Hence, for each tweet associated to a given profile, the authors counted the number of words belonging to each of the 10 categories. Each tweet was then lower-cased, stripped from punctuation, and tokenized. Any tweet containing the mention symbol (@), or that matched either the MySQL 5.62 list of stopwords[1], or Twitter specific stop words such as *rt* (retweet), was discarded. After this pre-processing stage, the authors calculated, as stated in Equation 2.17, for each user profile, the normalized fraction of each one of the aforementioned categories' c counts.

$$f_c = \frac{w_c - \mu_c}{\sigma_c} \tag{2.17}$$

In the previous equation, w_c denotes the fraction of words belonging to category c over the total number of classified words, for a given profile, while μ_c corresponds to this same fraction averaged across all profiles, and σ_c corresponds to its standart deviation.

Regarding the classification into types, the authors considered five types of users: (1) influential, (2) popular, (3) listener, (4) star, and (5) highly-read. Cha *et al.* (2010) and Romero *et al.* (2010) independently found that user's influence is related with the amount of mentions and retweets that a given user suffers, and that user popularity highly correlates with the number of followers. Based on this previous knowledge, and in order to classify a user as influential, the authors relied on two different tools: Klout[2] and TrstRank[3]. Klout measures to what extent contents from a given user are replied and retweeted, while TrstRank measures how important a user is, based

[1] http://dev.mysql.com/doc/refman/5.6/en/fulltext-stopwords.html
[2] http://klout.com
[3] http://api.infochimps.com/describe/soc/net/tw/trstrank

on more complex measures such as the betweness centrality of the user in the Twitter graph. For identifying the other three types of users, the authors leverage on the statistics already available in the collected Twitter dataset (i.e., number of followers, followees, and the number of users that added that given user to their reading list). A given user is classified as (1) listener if the number of users he follows exceeds a given threshold, as (2) star if his followers to followees ratio exceeds a given threshold, and as (3) highly-read if the number of times he is listed in other users' reading list exceeds a given threshold.

For studying the relationship between the aforementioned ten linguistic categories from the LIWC lexicon, and the five types of users, the authors compute the correlation coefficients of a linear regression model between the normalized countings of the ten LIWC categories and the logarithm of each of the measures' values associated to each one of the user types (e.g., followers to followees ratio for the stars). The authors only considered correlation coefficients whose p-value < 0.001. From the considered coefficients, the authors collected some insights: (1) listeners, popular users and highly-read users all use language that promotes one-to-one engagement, (2) influential users resort to language that creates a sense of community, and (3) stars usually use self-centered language. Concerning categories related to emotions (i.e., the aforementioned *posemo* and *negemo* categories), the authors also found that (4) popular users mainly use positive emotions, while (5) influential users express primarily negative emotions.

As a last insight into the relationship between emotional expression through the use of language and influence, the authors presented a sentiment metric, inspired in the work of Kramer (2010). This metric, presented in Equation 2.18, returns, given a user i, an emotional score associated to the messages he produced.

$$\text{Sentiment}(i) = \frac{p_i - \mu_p}{\sigma_p} - \frac{n_i - \mu_n}{\sigma_n} \tag{2.18}$$

In this Equation, p_i and n_i denote, respectively, the fraction of positive and negative words associated to the user i, and μ_p and μ_n denote these same fractions averaged across all users. Finally, σ_p and σ_n correspond to the standard deviation of the aforementioned fractions across all users. This normalization, using the mean and the standard deviation, intends to counteract the unbalanced distribution of positive and negative words in the English language. Then, the authors measured the correlation between this measure and influence (i.e., using the Klout and Trstrank scores), and found a strong negative correlation (i.e., a correlation coefficient of -0.924 and a correlation coefficient of -0.599, when respectively using the Klout and Trstrank scores). The authors argue that these results suggest that users, in Twitter, are influenced by those who primarily express negative emotions.

2.2.2 Analysis of Social Media Content Propagation

This section presents previous work that studied how social media contents spread in networks, particularly focusing on studies that attempted to analyze spatio-temporal properties of information propagation.

Guille *et al.* (2013) presented an exhaustive overview of existent works around information diffusion in social networks, and proposed a taxonomy that summarizes the state-of-the-art methods and techniques dealing with this issue. The tutorial presented by Leskovec (2011) also describes techniques that attempt to address a set of problems that range from *extracting temporal patterns by which information popularity grows and fades over time* to *build predictive models of information diffusion and adoption*.

Spatial Influence Versus Community Influence

Kamath *et al.* (2012) examined the spread of social media content around the globe, trying to understand and to answer questions such as: (i) in what way does content spread from location to location, or (ii) can a model about this spread be built and used to predict the next ocurrences of a given content. A probabilistic model was developed, combining two opposed suppositions:

- A spatial influence assumption, where it is assumed that content is usually spread first to nearby locations;
- A community influence assumption, where it is assumed that content spreads over locations that have some cultural background in common. This second assumption is based on the fact that online communication is not subject to geospatial limitations. There are various methods to evaluate community affinity and the authors considered two options: (i) two communities are closely related if they have identical contents in common, independently of when they adopt these contents, or (ii) the two communities are close if they share identical contents at the same time.

In order to assess the importance of geo-spatial properties in the spread of social media contents, the authors considered the posting of hashtags on Twitter. They collected around 755 million geo-referenced hashtags, corresponding to about 10 million different hashtags. This sample was

pulled from Twitter's streaming API, in the period between February 1, 2011 and November 30, 2011. All tweets were transformed into a tuple: $< hashtag, time, latitude, longitude >$. Every different hashtag is taken into account as a single activity m. The set of all unique hashtags corresponds to the full set of activities M.

The geo-spatial properties that were measured by the authors are: (i) hashtag sharing versus distance, (ii) hashtag adoption lag versus distance, and (iii) predictability of hashtag spread.

When considering hashtag sharing versus distance, the authors aimed to find if distance has an impact in the amount of shared hashtags between two locations. The physical distance between two locations is calculated with the Haversine distance approach. The hashtag similarity of two locations is calculated using the formula in Equation 2.19, where l stands for a location and M_l for the set of different hashtags that occurred in l.

$$\mathrm{sim}_{\mathrm{hashtag}}(l_1, l_2) = \frac{|M_{l1} \cap M_{l2}|}{|M_{l1} \cup M_{l2}|} \tag{2.19}$$

The authors then explored the relationship between hashtag similarity and distance. This relationship has a strong correlation. The closer two places are, the more similar they are. Albeit the correlation being high, there are still some outlier locations, which the authors argue that may be caused by cultural factors.

The measurement of hashtag adoption lag versus distance reflects how synchronized was the adoption of a common hastag related to the physical distance between two locations. The hashtag adoption lag is defined as shown in Formula 2.20, where t_l^m is the time activity m first appeared in location l.

$$\mathrm{lag}_{\mathrm{adoption}}(l_1, l_2) = \frac{1}{|M_{l1} \cap M_{l2}|} \sum_{m \in M_{l1} \cap M_{l2}} |t_{l1}^m - t_{l2}^m| \tag{2.20}$$

The authors also found a correlation between hashtag adoption lag and distance. Locations physically closer tend to adopt hashtags at the same time. As in the case of hashtag sharing, there are some outliers cases that the authors assign again to possible cultural factors.

The calculation of the predictability of spread makes usage of coverage, as defined in the following equation:

$$\mathrm{C}(O^m) = \frac{1}{|O^m|} \sum_{o \in O^m} \mathrm{D}(o, \mathrm{G}(O^m)) \tag{2.21}$$

Equation 2.21 essentially corresponds to the mean distance from all the occurrences of an activity

m to its midpoint. In the formula, D is the Haversine distance function and G is the geographic midpoint. The parameter O^m refers to the set of all observations of the activity m.

To figure out the predictability of spread, the authors compared the coverage of hashtags after their full propagation, and their coverage after a smaller time interval. They found that the final coverage of a hashtag can be predicted with a good certainty just after some minutes after its initial occurrence. This predictability raises when the innitial time interval increases.

To answer to the question on if it is possible to predict where a content will reach, the authors attempted to model the hashtag spread. They formalized the question with what they call the location subset selection problem, which can be expressed through Equation 2.22. This formalization returns the k top locations that have the most amount of unobserved occurrences of a hashtag m at a time t_s. The parameter U_S^m corresponds to the set of unobserved occurrences of hashtag m in the subset S of all locations L.

$$\mathrm{M}(m, L) = S_{ts}^m = \underset{S \subseteq L}{\operatorname{argmax}} U_S^m \tag{2.22}$$

This model can support the intuition that locations influence one another, depending on the distance among them, or depending on cultural factors. This influence may be abstracted by $I^{l_i \to l_j}$ which within a range of $[0, 1]$ conveys the influence that location i has in location j. Given an activity (i.e. a hashtag) m, a spread model for an influence $I^{l_i \to l_j}$ is given by:

$$\mathrm{M}_{\mathrm{spread}}(m, L) = S_{ts}^m = \underset{S \subseteq L}{\operatorname{argmax}} \sum_{l \in S} \left(\frac{O_{li}^m}{O^m} + \sum_{l_i \in L - l} \frac{O_{li}^m}{O^m} I^{l_i \to l_j} \right) \tag{2.23}$$

In the formula, $\frac{O_{li}^m}{O^m}$ is the estimated probability of observing m in l, since it is based on the propagation of m until t_s. The independence of this model on what regards the type of influence allowed the authors to define $I^{l_i \to l_j}$ in several ways: (i) spatial influence, that takes into account physical distance; (ii) transmitting influence, that considers temporal proximity; or (iii) sharing influence, that considers hashtag sharing. The authors also mixed different influence assumptions in the same model, giving different weights to each one of them.

Regarding the evaluation of the proposed model, the authors compared it against: (i) a random selection of locations, (ii) a greedy selection, that considers that the hashtags will continue to be diffused in the locations where they already are popular, or (iii) a selection based on a linear regression model. The authors found that their model, when combining community influence with spatial influence, has the best results. Specifically, the authors measured an accuracy of 42% when selecting the top k locations where an activity will be popular.

Spatio-Temporal Dynamics of Online Memes

Kamath *et al.* (2013) analyzed the diffusion of hashtags in Twitter, studying the effect that location, time and distance have in the diffusion process. This analysis was based on a collection of two billion geo-referenced tweets. These tweets contained 342 million hashtags, 27 million of them being unique, and they were obtained through Twitter's streaming API. The dataset was collected in the period between January 1, 2001 and October 31, 2012. Every tweet was transformed into a tuple of the form $< hashtag, time, latitude, longitude >$.

The spatial diffusion was analyzed through measures of focus, entropy and spread. All these metrics are based on probabilities computed through Equation 2.24, where h stands for a hashtag, l for a location, L for all possible locations, and O_l^h refers to the number of observations of a hashtag h in a location l. Hereafter in this presentation, P_l^h means the probability of observing a given hashtag h in a given location l.

$$P_l^h = \frac{O_l^h}{\sum_{l \in L} O_l^h} \tag{2.24}$$

The focus metric from Equation 2.25 corresponds to the maximum probability of observing a hashtag in one location only. Intuitively, the focus will diminish as propagation increases.

$$F^h = \max_{l \in L} P_l^h \tag{2.25}$$

By entropy, as shown in Equation 2.26, the authors mean the randomness of a hashtag's spatial distribution. A hashtag that only occurs in one location has an entropy of zero. When an hahstag propagates to other locations, its entropy will naturally increase.

$$E^h = -\sum_{l \in L} P_l^h \log_2(P_l^h) \tag{2.26}$$

The last measure, shown in Equation 2.27 and named spread, is the mean distance from all the occurrences of a hashtag to its midpoint. In the formula, D is the Haversine distance function, and G is the geographic midpoint. The parameter O^h refers to the set of all observations of a hashtag h. This metric is similar to the notion of coverage shown in Equation 2.21.

$$S^h = \frac{1}{|O^h|} \sum_{o \in O^h} \mathrm{D}(o, \mathrm{G}(O^h)) \tag{2.27}$$

By analyzing the Twitter dataset, the authors intended to know:

- How distance influences hashtag adoption between users;
- If hashtags really are a global phenomena, as it is usually said;
- What are the geo-spatial properties of hashtag spread;
- How global and local hashtags vary in relation to geo-spatial properties;
- How long does it take a hashtag to reach its peak in terms of being spread by many users, and how do geo-spatial properties influence this time;
- To what degree can locations be classified through the geo-spatial characteristics of the hashtags that they originated;
- What are the most influential locations.

To answer some of the aforementioned questions, the authors considered an approach, previously shown in Equation 2.19, to measure the similarity between pairs of locations, based on the fraction of hashtags shared among the two locations.

The analysis of the data has shown that physical distance, calculated using the Haversine approach, influences the quantity of shared hashtags among two locations. The authors observed a correlation between these two factors, and this suggests that the closer locations are, the greater the probability of adopting the same hashtags, and vice-versa. This may be explained, in great part, because of issues of culture, language and shared interests. Locations that are physically close are also found to adopt more hashtags at the same time and vice-versa.

When focus was being measured, the authors noticed that about one quarter of all hashtags are observed in a single location only. Hashtags with low occurrences tend to have a higher focus, meaning that low intensity hashtags manifest mainly in a single location. An increase of occurrences leads to a decreased focus. Based on the results of this measure the authors also suggest that a significant part of the hashtags are linked to local events and to conversations between friends that are located geographically close.

The measurement of entropy again reveals that about twenty-five percent of the hashtags are located in a single location, and that most hashtags propagate to at most two locations. It was also concluded that hashtags that occur many times are more likely to spread over several locations.

Observing hashtags through the lenses of spread can also reveal that about a quarter of them have a spread of zero, since they were observed in a single location. About half of the hashtags hold a spread inferior to four hundred miles, and only the remaining percentage has a spread

bigger than one thousand miles. A rising number of occurrences is linked to a higher spatial diffusion.

A direct comparison of these spatial properties leads to some interesting outcomes. As expected, an increase in spread results in a reduction in focus, as well as an increase in entropy. It is also possible to extract three major groups of hashtags: (i) those of local interest that have high focus, low entropy and have less than seven hundred miles of spread; (ii) those of regional interest and that are event motivated, having between seven hundred and one thousand miles of spread; and (iii) those that are hashtags related to global events.

Another aspect that is worthwhile to mention is the peak analysis of hashtag diffusion. In what concerns these peaks, hashtags may be split in two categories: (1) hashtags that reach their usage peak in about thirty minutes, where local hashtags peak faster than global, and (2) slow peaking hashtags, which reach their usage peak in between four and ten hours, and where global hashtags tend to peak relatively faster than local ones.

The authors also attempted to induce hashtag propagation patterns. The authors found that hashtags do not spread uniformly around their origin, but instead in steps, from city to city. It was also found that most hashtags receive most of their occurrences from a single location during their peak phase. Lastly, most hashtags are initially spread by a single location, which diffuses them over other locations. Even after the hashtag becomes popular, this original location continues to spread it.

Cross-Lingual Study on the Relation between Emotions and Virality

Guerini & Staiano (2015) studied the relation between the virality of news articles and the emotions these articles are found to raise on the readers. The authors aim at understanding if viral phenomena are consistently influenced by emotions, and verifying if these influences remain true across different languages (i.e., English and Italian). In order to assess if these hypothesis are according to the reality, the authors leverage on two datasets: (1) a set composed of 53.226 news from the website `www.rappler.com`, in which every displayed page contains an interface where the users are asked to select the emotion the news article evokes on them (e.g., *happy*, *sad*, *angry*, etc.) and (2) a set of 12.437 news articles crawled from the online version of the newspaper *Corriere della Serra*, which has adopted an identical approach, based on *smileys*, to register the emotions sensed by the readers when consuming a given content.

The authors split viral phenomena they study into two distinct categories: (1) narrow-casting and (2) broadcasting. *Narrow-casting* consists in diffusing a content to a restricted audience, i.e.,

for the purpose of this work, commenting a given news article, in the *comment box area* of the page. Analogously, the authors interpret *broadcasting* as sharing a given article into the social networks.

So that the relation between emotional labels assigned by the readers to the articles and the virality indices may be studied (i.e., comments, threads, g+ shares, Facebook shares, Twitter shares), the authors resorted to simple linear models. When looking into the results obtained when building the models with the dataset from `www.rappler.com`, on one hand the authors found an influence from emotions like *inspiring* and *anger* into the virality of a content, as on the other hand they found that *sadness* is related with contents with low virality. They found no significant difference in these relations when considering narrow and broadcasting. When the authors repeated this same analysis with the dataset from *Corriere della Serra*, and when looking into narrow-casting the results remained relatively static in relation to the experiment with the previous dataset. However, when focusing on broadcasting they found *sadness* to have a major role in virality. The authors suggest that may be valid to hypothesize that cultural differences arise when acting upon a given emotion.

The authors then tried to measure these same relations but through the lenses of the VAD (valence, arousal and dominance) emotional model that maps emotions into a 3-dimensional space (Russel, 1980). Thus, the authors mapped the emotions that each one of the articles was found to evoke to a 3-dimensional vector. They made this conversion by multiplying the percentage of votes for each emotion in a given article by the VAD scores assigned by Warriner *et al.* (2013b) to this same emotion. The results when assessing the relation between these scores and the virality indices show interesting resuls, since VAD dimensions are found to be consistent between the two considered datasets in both broadcasting (tweets/g+ shares) and narrowcasting (comments). These findings suggest that the aforementioned cultural divergences, when measuring the relation between emotions and virality, vanishes when taking into account their deeper VAD emotional constituents. The authors also find that users tend to narrow-cast articles which evoke a high arousal but a small dominance. Nonetheless, users tend to broadcast when they feel more in control, i.e., when the evoked dominance is higher. Finally, the valence dimension is found to be consistently related to all the indices of virality in both datasets, i.e., low valence is linked to a higher virality.

2.2.3 Overview

This section aims at making an overview of the work described in the present section, focusing on describing its major limitations, and how it relates with the work described in the following chapters.

One key aspect to mention regarding the works described in Section 2.2.1, which are related with the extraction of emotions from social network contents, is that some of these approaches only characterize the emotions in a simplified manner, i.e., measure how positive/happy or negative/unhappy a content is (Dodds *et al.*, 2011b; Li *et al.*, 2014; Quercia *et al.*, 2011). Other approaches, while characterizing emotions in a richer form, i.e., in terms of valence, arousal, and dominance, rely on naive methods for inferring the emotions encoded in a content: they resort to techniques that consist in counting words from emotional lexicons (Loff *et al.*, 2015). The method proposed in this dissertation addresses this set of limitations, by characterizing emotions in terms of valence, arousal, and dominance, as well as leveraging on regression models that based on neural embeddings of words and short texts predicts emotional scores in the aforementioned emotional dimensions.

The studies described in Section 2.2.2, are mainly related with the study on the diffusion of contents from social media, namely the spatio-temporal dynamics of their diffusion. Making exception to Guerini & Staiano (2015) where the authors try to study the relation between the emotional content of news, and how the news spread, most of these studies try to study the diffusion of contents without looking into the characteristics of these same contents. The second application presented in Chapter 5 goes in line with the work from Guerini & Staiano (2015), i.e., in the aforementioned application is tried to see if there is a correlation between emotional scores extracted from tweets, and the way these same tweets spread (e.g., the number of users they reach, the geographical coverage, etc).

Chapter 3

Predicting Affective Norms for Words

Lexical norms related to the emotional responses evoked by particular words, such as their valence, arousal, and dominance, are important research resources for many different fields. However, collecting such norms by asking human judges to rate sets of words is both expensive and time consuming, which strongly limits the size and availability of lexicons with emotional norms. In this Chapter, a technique for estimating lexical norms automatically is proposed by leveraging unsupervised word embeddings obtained through modern approaches based on neural networks, together with state-of-the-art approaches for building predictive models that leverage these embeddings as features.

Moreover, is also proposed a method based on Canonical Correlation Analysis, inspired on previous work by Faruqui & Dyer (2014) for leveraging English data with the purpose of estimating emotional lexical norms for other languages (i.e., Spanish, Portuguese, Italian and German). The obtained experimental results attest to the effectiveness of the proposed approaches.

These methods surpass the current state-of-the-art when predicting emotional ratings for English words, and in cross-language experiments, was also measured a significant correlation with human-ratings.

3.1 Introduction

Human emotional ratings of valence, arousal, and dominance, for particular words, are nowadays frequently used within cognitive science, behavioral psychology and psycholinguistic research, e.g. to study the cognitive mechanisms of emotional attention, word recognition, and numerous other phenomena in which emotions are hypothesized to play a key role. More recently, emotion ratings for words have also started to be used in a variety of studies leveraging text mining approaches over very large repositories, due to their utility in investigating a wide range of topics – see for instance the Hedonometer[1] project concerned with the usage of social-media data for measuring the happiness of large populations in near real time (Dodds *et al.*, 2011b), or the World-Well Being Project[2] concerned with studying psychosocial phenomena through language analysis, in which researchers have for instance shown that language patterns reflecting negative emotions on social media messages, especially anger, are strong markers of cardiovascular mortality (Eichstaedt *et al.*, 2015).

Emotional ratings are typically collected through interviews, by asking participants to rate words according to the emotional dimensions under consideration. A dimension of valence can be defined as the pleasantness of the stimulus, and it can for instance be operationalized by asking participants to rate how they feel while reading the word, on a scale from one (*happy, pleased, satisfied, contented* or *hopeful*) to nine (*unhappy, annoyed, unsatisfied, melancholic, despaired* or *bored*). Arousal can, in turn, be identified with the intensity of feeling being evoked by a particular word, and it can be rated on a scale from *stimulated, excited, frenzied, jittery, wide-awake* or *aroused* to *relaxed, calm, sluggish, dull, sleepy* or *unaroused*. Finally, a dimension of dominance can be identified with the degree to which the word makes the reader feel *in control*, and participants can rate their emotions while reading the word on a scale from *in control, influential, important, dominant, autonomous* or *controlling* to *controlled, influenced, cared-for, awed, submissive* or *guided*.

Notice, nonetheless, that collecting emotion norms from human raters is both expensive and time consuming. As a result, affective norms are only available for a few English words, are not available for proper nouns even in English (Recchiaa & Louwerse, 2014), and are sparse in other languages. The Affective Norms for English Words (ANEW) dataset remains the most frequently used set of emotion norms (Bradley & Lang, 1999). The original ANEW dataset consists of 1,034 words rated according to valence, arousal and dominance, with a 2010 update bringing the total up to 2,471 unique words (Bradley & Lang, 2010). Despite the recent introduction of a set of norms of valence, arousal and dominance for 13,915 English lemmas (Warriner *et al.*,

[1] http://hedonometer.org/
[2] http://wwbp.org

2013a), or despite similar crowd-sourcing efforts being reported in the literature (Dodds *et al.*, 2011b), many vocabulary terms are still not covered by existing collections of norms, and large-scale emotion ratings are hard to come by for many languages (e.g., the largest sets of norms for Spanish (Redondo *et al.*, 2007), Portuguese (Soares *et al.*, 2012), Italian (Montefinese *et al.*, 2014) and German (Schmidtke *et al.*, 2014) respectively consist of 1,034, 1,034, 1,121 and 1,003 words, in an attempt to replicate those from the original ANEW study).

The aforementioned limitations have motivated researchers to seek automated procedures for estimating affective norms, for instance through word co-occurrence patterns leveraging Latent Semantic Analysis (Bestgen & Vincze, 2012) or Point-wise Mutual Information statistics, and/or through regression modeling approaches (Manderaa *et al.*, 2015). Being able to automatically construct or extend emotional norms would for instance offer new perspectives for optimizing word selection in factorial experiments, and for drawing large samples for multiple-regression studies. If valence, arousal and dominance ratings, as estimated automatically from auxiliary data, can be found to correlate well with human ratings (e.g., if the measured correlations are comparable to previously reported correlations among sets of human ratings), then one such automated procedure can be used to quickly and inexpensively generate approximate ratings for any previously unseen word. This may, for instance, have important applications in the development of algorithms for estimating emotional norms for entire sentences or texts (Calvo & D'Mello, 2010; Paltoglou & Thelwall, 2013; Paltoglou *et al.*, 2013).

In recent years, several unsupervised methods based on neural network architectures have been proposed to derive word embeddings from large corpora. In this context, word embeddings correspond to dense vector representations that implicitly capture syntactic and semantic properties of words (i.e., we have that a notion of semantic similarity, as well as other linguistic regularities, seem to be encoded in the embedding spaces resulting from these methods (Mikolov *et al.*, 2013)). Word representations based on neural network architectures have been shown to outperform other distributional similarity approaches (Baroni *et al.*, 2014), and is a goal of this work to argue that these embedding vectors can be used as features to train a model for predicting the emotional properties for new words. In this thesis, following the ideas of Recchiaa & Louwerse (2014), this particular hypothesis, was tested.

Taking inspiration on recent developments within computational linguistics, was also explored the possibility of making cross-language extrapolations for psycholinguistic variables. It was recently shown that it is possible to learn a linear mapping between vector spaces of two languages (Mikolov *et al.*, 2013), and authors such as Faruqui & Dyer (2014) have argued that lexicon-semantic content should be invariant across languages, proposing techniques (e.g., based on canonical correlations analysis) for improving unsupervised word embeddings, generated

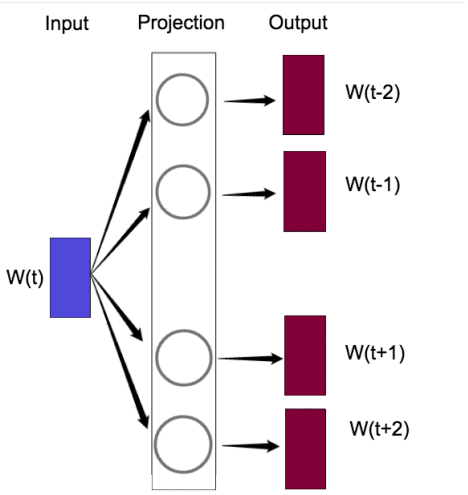

Figure 3.3: The skip-ngram model from word2vec for learning word embeddings (Mikolov *et al.*, 2013), where the training objective is to learn word vector representations that are good at predicting nearby words.

mono-lingually, through the incorporation of multilingual evidence. In addition to estimating norms in a given language, was also attempted to see if information from other languages can be used when extrapolating ratings (e.g., were used sets of ratings that were already collected for English, together with word embeddings in both English and a target language after re-projection into a common vector space, to predict ratings for a target language).

3.2 Neural Word Embeddings

Word embeddings are generally trained by optimizing an objective function that can be measured without annotations. One popular objective is to maximize the prediction of contextual words. In the work described by Mikolov *et al.* (2013), commonly referred as word2vec's skip-ngram model, the idea is to estimate the optimal word embeddings by maximizing the probability that the words within a given window size are predicted correctly, leveraging a simple two-layer neural network in which the top-layer corresponds to a log-linear model – see Figure 3.3. Given the i-th word from a sentence w_i, the skip-n-gram approach models the probability of each word at a distance p from w_i, according to:

$$\mathrm{p}(\mathbf{w}_{i+p}|\mathbf{w}_i, \mathbf{C}_p, \mathbf{E}) \propto \exp(\mathbf{C}_p \cdot \mathbf{E} \cdot \mathbf{w}_i) \tag{3.28}$$

In the formula, $\mathbf{w}_i \in \{1, 0\}^{v \times 1}$ is a one-hot representation of the word (i.e., a sparse column vector

of the size of the vocabulary v, with a one on the position corresponding to the word). The symbol $\cdot$ denotes the dot product, and $\exp()$ acts element-wise. The log-linear model is parametrized by two matrices $\mathbf{E}$ and $\mathbf{C}_p$. The matrix $\mathbf{E} \in \mathbb{R}^{e \times v}$ is the embedding matrix, transforming the one-hot sparse representation into a compact real valued space of size e. Finally, $\mathbf{C}_p \in \mathbb{R}^{v \times e}$ is a matrix mapping the real-valued representation to a vector with the size of the vocabulary v. A distribution over all possible words is then attained by exponentiating and normalizing over the v possible options. In practice, due to the large size of v, various techniques are used to avoid having to normalize over the whole vocabulary. Some solutions proposed to address this problem include the usage of a hierarchical softmax objective function, or resorting to negative sampling (Goldberg & Levy, 2013; Mikolov *et al.*, 2013).

Stochastic gradient descent, computed using a back-propagation rule, is used to learn the parameters of these models, leveraging errors made in the predictions to adjust the parameters accordingly. After training, the low dimensional embedding $\mathbf{E} \cdot \mathbf{w}_i \in \mathbb{R}^{e \times 1}$ encapsulates the information about each word $\mathbf{w}_i$ and its surrounding contexts. For more information about word2vec's skip-ngram model, please refer to the original paper by Mikolov *et al.* (2013), or to the additional explanations provided by Goldberg & Levy (2013). In this work, were used 300-dimensional word embeddings, pre-trained on a Google News dataset that contains approximately 100 billion words, originally made available in word2vec's website[1].

3.3 Experiments in a Monolingual Setting

The experiments began by experimenting with the usage of English words in the set of norms from Warriner *et al.* (2013a) that did not appear in the ANEW corpus, as training data for predictive models that can later be used to estimate valence, arousal and dominance ratings for previously unseen words. Word embeddings were leveraged as features within different types of regression approaches, and evaluate the obtained results in the task of predicting the valence, arousal and dominance ratings in ANEW.

The aforementioned word representations were used together with three different types of forecasting models, namely a k nearest neighbor interpolation approach, random forest regression, and kernel ridge regression. The three approaches were implemented through the scikit-learn library (Pedregosa *et al.*, 2011), and they all naturally apply to multi-output problems, where the same predictor variables are used to predict several outputs (i.e., in this case, valence, arousal and dominance scores are predicted simultaneously).

[1] `https://code.google.com/p/word2vec/`

	ANEW					
	Valence		Arousal		Dominance	
	Pearson	MAE	Pearson	MAE	Pearson	MAE
k-NN	0.869	0.944	0.673	0.876	0.731	**0.551**
Random Forest	0.787	1.303	0.518	1.014	0.663	0.643
Kernel Ridge	**0.908**	**0.715**	**0.738**	**0.804**	**0.757**	0.560

	Warriner et al.					
	Valence		Arousal		Dominance	
	Pearson	MAE	Pearson	MAE	Pearson	MAE
k-NN	0.895	0.821	0.714	0.596	0.838	0.558
Random Forest	0.817	1.214	0.559	0.750	0.748	0.745
Kernel Ridge	**0.934**	**0.575**	**0.769**	**0.527**	**0.856**	**0.480**

Table 3.1: Obtained results when predicting ratings for words in the English ANEW lexicon Bradley & Lang (1999) and in the lexicon from Warriner *et al.* (2013a). The associated p-values for the Pearson product-moment correlation coefficient were always lower than 0.001.

In the k nearest neighbor interpolation approach, for each word in the test set, we identify the set of k most similar words (as measured according to the Euclidean distance between the word embeddings) in the training set, and assign the weighted mean rating of these words to the target word, as the extrapolated rating. The k nearest neighbors are weighted such that nearby instances contribute more to the final scores than faraway instances, namely by considering weights proportional to the inverse of the distance from the query instance. The value for k is an optimization parameter associated to this method.

Random forests are a general-purpose machine-learning technique based on an ensemble of randomized decision trees (Breiman, 2001). This method is based on building a set of decision trees, where each tree is based on a slightly different sample of the full dataset, reducing the risk of over-fitting the model. Each tree in the ensemble is built through the CART algorithm (Breiman *et al.*, 1984) from a sample drawn with replacement from the entire training set (i.e., a bootstrap sample). In addition, when splitting a node during the construction of the tree, the split that is chosen is the best split among a random subset of the features. As a result of this randomness, the bias of the forest usually slightly increases, with respect to the bias of a single non-random tree. However, due to averaging, the variance also decreases, usually more than compensating for the increase in bias and hence yielding overall better models. The CART algorithm constructs each binary decision tree in the ensemble using the feature and threshold that yield the lowest mean-squared error at each node. Given our multi-output setting (i.e., is simultaneously attempted to predict valence, arousal and dominance), the leaves of the trees store three output values, and the splitting criteria computes the average mean-squared error across all three outputs.

The main parameters to adjust when using random forests correspond to the number of trees in

the forest (i.e., the larger the better, but also the longer it will take to compute) and the size of the random subsets of features to consider when splitting a node (i.e., the lower the greater the reduction of variance, but also the greater the increase in bias). An empirically good approach for the case of regression problems is to set the size of the random subset of features equal to the total number of features. As for the number of trees, it was fixed at 300 in these experiments.

Finally, the kernel ridge regression approach combines the standard ridge regression (i.e., linear least squares with l2-norm regularization) with the kernel trick, as used in Support Vector Machines. It thus learns a linear function in the space induced by the respective kernel and the data, which for non-linear kernels corresponds to a non-linear function in the original space. Standard ridge coefficients minimize a penalized residual sum of squares, corresponding to:

$$\min_{\mathbf{w}} ||\mathbf{X} \cdot \mathbf{w} - \mathbf{y}||_2^2 + \alpha||\mathbf{w}||_2^2 \tag{3.29}$$

In the previous formula, $\mathbf{X}$ is the matrix of explanatory variables, $\mathbf{y}$ is a vector with the target values, and $\alpha \geq 0$ is a regularization parameter that controls the amount of shrinkage (i.e., the larger the value of α, the greater the amount of shrinkage, and thus the coefficients become more robust to co-linearity). In our case, given that the aim is predicting valence, arousal and dominance, independent ridge regression models were built, i.e. one for each of the three outputs. Kernel ridge regression extends the general setup considered above to allow for nonlinear prediction functions. For an arbitrary instance $\mathbf{x} \in \mathbb{R}^n$, the outcome suggested by ridge regression (i.e., $\mathbf{w}^\intercal \cdot \mathbf{x}$) can be rewritten into the dual form of the ridge regression solution (i.e., $\mathbf{w}^\intercal \cdot \mathbf{x} = \mathbf{y}^\intercal(\alpha\mathbf{I} + \mathbf{X}\mathbf{X}^\intercal)^{-1}\mathbf{X}\mathbf{x}$). When using the dual form, and because we only need scalar products between instances, we can directly use a kernel function to map instances into a higher-dimensional feature space, where regression can often be made more effectively. A popular choice for the kernel, which were used in our experiments, is a radial basis function of the form $\mathrm{k}(\mathbf{x}_a, \mathbf{x}_b) = \exp(-\gamma|\mathbf{x}_a - \mathbf{x}_b|^2)$ with $\gamma > 0$. The values for α and γ are optimization parameters associated to the kernel ridge regression method.

Table 3.1 presents the results obtained in our first set of experiments, both in terms of Pearson's correlation coefficient r, and in terms of the Mean Absolute Error (MAE). These metrics can be computed as shown in the equations below, where x and y are sets with the obtained results and the ground truth measurements, and where $|e_i|$ is the absolute error for a testing instance i.

$$\mathrm{r}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.30}$$

$$\mathrm{MAE}(x, y) = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i| \,. \tag{3.31}$$

The parameters associated to the k nearest neighbor and kernel ridge regression approaches were tuned through a simple grid-search, so as to optimize the average scores in all three emotional dimensions. By optimizing parameters according to the average correlation scores, overfitting the models to individual cases, is avoided. The best results were obtained for $k = 19$, $\alpha = 0.1$ and $\gamma = 1$. A total of 12,764 words were used for model training, and evaluation was mostly made through the 1,026 words present in the ANEW lexicon. Nonetheless, are also presented results when considering ratings for these same 1,026 words, as available in the dataset by Warriner *et al.* (2013a). All the rated words were also present in the dataset of pre-trained word embeddings given at word2vec's website.

The obtained results attest to the effectiveness of the proposed method, as were obtained very high correlations by leveraging the representations based on skip-ngram embeddings, even without considering additional features as done in the study by Recchiaa & Louwerse (2014). The results obtained with the three different types of prediction models are relatively similar, although the kernel ridge regression approach outperformed the others in terms of Pearson's correlation, over all 3 emotional dimensions. These correlation values are similar to those reported on the previous studies by Bestgen & Vincze (2012) and by Recchiaa & Louwerse (2014), even slightly superior. Still, details of the extrapolation procedures in these studies are too heterogeneous to allow for a direct comparison of their efficiency (i.e., the authors used different sets of predictors, information derived from different corpora, different kinds of models, and different validation procedures).

For comparison, correlations between the valence, arousal and dominance ratings given in the original ANEW, by Bradley & Lang (1999) and in the study by Warriner *et al.* (2013a) are, respectively, of 0.953, 0.761 and 0.795. In the study by Warriner *et al.* (2013a), the authors report that typical correlations of human ratings across languages range from 0.85 to 0.97 for valence, 0.56 to 0.76 for arousal, and 0.77 to 0.83 for dominance, whereas when considering the English language, split-half reliabilities across human participants are 0.91 for valence, 0.69 for arousal, and 0.77 for dominance. Correlations are somewhat lower between English speakers of different genders (i.e., 0.79, 0.52 and 0.59, for valence, arousal and dominance), different ages (i.e., correlations of 0.82, 0.50 and 0.59 when comparing subjects younger than 30 versus older than 30, respectively for valence, arousal and dominance), and different educational backgrounds (0.83, 0.47 and 0.61, respectively for valence, arousal and dominance), but remain large overall. In general terms, the automatically estimated ratings obtained using the proposed method are at

least as correlated with human ratings as male/female, old/young, and high/low education English speakers' ratings are with each other, and in many cases even more so.

In a separate test, was also attempted to see if the errors produced by the proposed method correlate with the standard deviation observed in the ratings produced by data collection from human subjects. Therefore, the absolute difference between the estimated and ground-truth ratings, is calculated, afterwards measuring its correlation towards the standard deviation in the human ratings. Were obtained correlation results in the range of $[-0.07, 0.05]$, in all three emotional dimensions and for both the ANEW norms and those from Warriner *et al.* (2013a). These results suggest that there is no significant correlation between the errors generated by the automated procedures proposed here, and those cases where human subjects also present a higher variability.

3.4 Experiments in a Cross-lingual Setting

It was also attempted to use information from the English language for extrapolating ratings to other languages, specifically Portuguese, Spanish, Italian and German, later leveraging adaptations of the original ANEW dataset into these four separate languages in order to evaluate the proposed approach (Montefinese *et al.*, 2014; Redondo *et al.*, 2007; Schmidtke *et al.*, 2014; Soares *et al.*, 2012). Representations were used for the English words in the set of norms from Warriner *et al.* (2013a), specifically for words that do not appear in the ANEW corpora for each target language, as training data for the predictive models. The representations for the English words are based on the same 300-dimensional skip-ngram word embeddings that were pre-trained on a Google News dataset, originally made available in word2vec's website. However, in order to train predictive models that can later be used for extrapolating ratings to other languages, there is the need to represent words in the target language in the same embedding space as the training data. Taking inspiration on a previous work by Faruqui & Dyer (2014), canonical correlations analysis was used to project two sets of word embeddings, trained separately for each language, into a common representation space. Monolingual skip-ngram word embeddings for the Portuguese, Spanish, Italian and German languages were first trained with the word2vec software tool, leveraging recent Wikipedia dumps for these four languages. The dimensionality of the word embeddings was set at 300, and was used one training epoch over each of these datasets, together with a contextual window size of 10 words.

In general, we have that Canonical Correlations Analysis (CCA) (see Fig. 3.4) computes linear

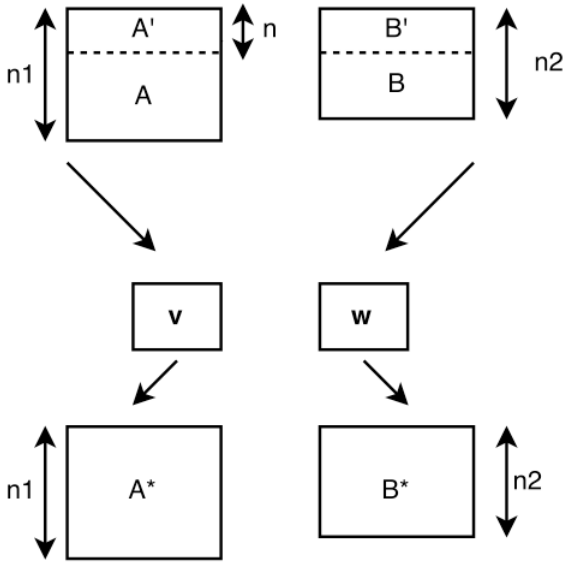

Figure 3.4: Projection of cross-lingual word embeddings using Canonical Correlations Analysis (CCA).

transformations of a pair of random variables, such that their projections are maximally correlated. Leveraging a seed set of translation word pairs, CCA can be used to transform the representations of these seed words, so as to maximize their correlation. Afterwards, the same linear transformations can be applied to all words in each language. The previous work by Faruqui & Dyer (2014) has shown that CCA can be used to build better word representations, by capturing multilingual evidence, different aspects of word meaning and different types of distributional profiles for the words. Another appealing property of CCA is that, if there is noise in either view that is uncorrelated with the other view, the learned joint representations should not contain the noise in the uncorrelated dimensions.

Let $\mathbf{A} \in \mathbb{R}^{n_1 \times d_1}$ and $\mathbf{B} \in \mathbb{R}^{n_2 \times d_2}$ be matrices corresponding to word embeddings of two different vocabularies, where the rows represent individual words. Since the two vocabularies are of different sizes (i.e., n_1 and n_2) and there might not exist translations for every word of $\mathbf{A}$ within $\mathbf{B}$, let $\mathbf{A}' \subseteq \mathbf{A}$ represent a subset of words, where every word in $\mathbf{A}'$ is translated to one other word in $\mathbf{B}' \subseteq \mathbf{B}$. We thus have that $\mathbf{A}' \in \mathbb{R}^{n \times d_1}$ and $\mathbf{B}' \in \mathbb{R}^{n \times d_2}$.

Let $\mathbf{x}$ and $\mathbf{y}$ be two corresponding vectors from $\mathbf{A}'$ and $\mathbf{B}'$, and let $\mathbf{v}$ and $\mathbf{w}$ be two projection

directions. Two projected vectors can be computed from $\mathbf{x}' = \mathbf{x} \cdot \mathbf{v}$ and $\mathbf{y}' = \mathbf{y} \cdot \mathbf{w}$, and the correlation between the projected vectors can be written as:

$$\rho(\mathbf{x}', \mathbf{y}') = \frac{\mathrm{E}[\mathbf{x}' \cdot \mathbf{y}']}{\sqrt{\mathrm{E}[\mathbf{x}']^2 \times \mathrm{E}[\mathbf{y}']^2}} \tag{3.32}$$

CCA maximizes the correlation ρ for the given sets of vectors $\mathbf{A}'$ and $\mathbf{B}'$, and it outputs two projection vectors $\mathbf{v}$ and $\mathbf{w}$ that result in this maximal correlation (i.e., $\mathbf{v}, \mathbf{w} = \mathrm{CCA}(\mathbf{x}, \mathbf{y}) = \arg\max_{\mathbf{v},\mathbf{w}} \rho(\mathbf{x} \cdot \mathbf{v}, \mathbf{y} \cdot \mathbf{w})$). Using the two resulting vectors $\mathbf{v} \in \mathbb{R}^d$ and $\mathbf{w} \in \mathbb{R}^d$, we can project the entire vocabulary of the two languages $\mathbf{A}$ and $\mathbf{B}$ through a simple dot product, and thus CCA solves the problem of not having translations of a particular word in the seed dictionary. Notice, however, that the dimensionality d of the resulting vectors cannot be longer than that of the original monolingual vectors (i.e., $d = \min(d_1, d_2)$).

In these experiments, were computed 300-dimensional word embeddings from bilingual projections between the English-Spanish, English-Portuguese, English-Italian, and English-German language pairs, leveraging the aforementioned pre-trained monolingual embeddings. The seed sets of translation word pairs, used for training the CCA projections, were obtained from UWN[1], a multilingual lexical knowledge base based on WordNet and Wikipedia (de Melo & Weikum, 2009, 2010). In UWN, words in multiple languages are associated to a corresponding list of meanings, and were therefore used pairs of words with a same meaning to automatically build the seed sets.

Forecasting models were trained with basis on the re-projected word embeddings for the English words in the study by Warriner *et al.* (2013a), and they were then applied to the re-projected embeddings for words in the Spanish (Redondo *et al.*, 2007), Portuguese (Soares *et al.*, 2012), Italian (Montefinese *et al.*, 2014) and German (Schmidtke *et al.*, 2014) versions of ANEW. The same forecasting models described in the last section were used in this particular set of experiments.

Table 3.2 presents the results obtained in the second set of experiments. The parameters associated to the k nearest neighbor and kernel ridge regression approaches were again tuned through a simple grid-search, so as to optimize the average correlation scores in all three emotional dimensions, and across the four languages. The best averaged results where obtained for $k = 91$, $\alpha = 0.1$ and $\gamma = 0.1$. Table 3.2 also shows the number of words used in the training and testing of each model, as well as the number of words in the seed set of translations for CCA. The words used for model training had to be present in the pre-trained embeddings provided in word2vec's

[1] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/uwn/

	Number of Words			k-**NN**		
	Training	Testing	Seeds	Valence	Arousal	Dominance
ES	12783	1030	18822	0.613	0.429	0.610
PT	12783	1009	13197	0.565	0.420	0.443
IT	12713	1111	20070	0.586	0.411	0.500
DE	12813	981	9555	0.470	0.332	0.393

	Random Forests			**Kernel Ridge**		
	Valence	Arousal	Dominance	Valence	Arousal	Dominance
ES	0.641	0.434	0.596	**0.674**	**0.445**	**0.642**
PT	0.583	**0.447**	0.483	**0.659**	0.391	**0.505**
IT	0.619	**0.454**	0.520	**0.637**	0.453	**0.527**
DE	**0.570**	**0.466**	**0.414**	0.567	0.411	0.402

Table 3.2: Pearson correlations obtained when predicting the ratings in four different adaptations of the ANEW lexicon, namely for the Spanish, Portuguese, Italian and German languages. The corresponding p-values were always lower than 0.001.

website, whereas the words used for model testing had to be present in the embeddings trained by us with the word2vec software and leveraging recent Wikipedia dumps, as well as in each respective adaptation of the ANEW norms.

The obtained results show that relatively high correlations can be achieved for all four languages, although they are much inferior to the results obtained for the monolingual setting. For comparison purposes, Table 3.3 shows the correlations between the norms for valence, arousal and dominance, in the original ANEW dataset and in the set provided by Warriner *et al.* (2013a), against the norms in the four different adaptations of the ANEW dataset. It is interesting to notice that a higher predictive accuracy is generally also obtained for the languages where the correlation towards the English norms is higher (i.e., Italian and Spanish).

Since CCA gives us projection vectors sorted in descending order of correlation, were also performed experiments by taking projections of the original word vectors across only the top n correlated dimensions. The parameters involved in the optimization of the predictive models were kept at the same values used in the experiments reported on Table 3.2 (i.e., $k = 91$, $\alpha = 0.1$ and $\gamma = 0.1$). Figures 3.5 and 3.6 show the obtained results for all four languages, and when varying the n parameter between $\{99, 150, 240, 300\}$. The correlation values remain relatively stable when varying this parameter, and it is interesting to notice that the results with lower-dimensional projections are even slightly superior to those obtained with the full-dimensional projected embeddings. For instance in the case of the Spanish ANEW, when using a dimensionality of $n = 99$, the Pearson correlations for valence, arousal and dominance are respectively of 0.718, 0.515 and 0.667. These findings are in agreement with the experiments reported by Faruqui & Dyer (2014), who also observed that better cross-lingual representations for words could be obtained

when taking only the top n most correlated dimensions, usually obtaining the best results when $n = 80\%$ of the original dimensionality.

For comparison purposes, was also experimented the training of kernel ridge regression models (i.e., the best performing method in the previous experiments) leveraging monolingual data (i.e., leveraging the skip-ngram embeddings trained separately for each of the four languages, together with the ANEW norms adapted to each of these languages), using a leave-one-out cross validation methodology for evaluating the quality of the obtained results. The parameters k, α and γ were again kept at the same values considered for the experiments reported on Table 3.2. Table 3.5 presents the results from this particular experiment, showing that the obtained correlations are relatively similar to those obtained with the CCA methodology. This finding further attests to the fact that the CCA methodology can be a useful alternative to derive lexicons of emotion ratings for languages where no such norms exist, given that the resulting estimates will likely have a similar quality to those that would be obtained by extrapolating from small amounts of data in the target language.

In a final test, was attempted to see if the size and quality of the set of seed translations influences the CCA projections that are obtained, and consequently the obtained predictions. Words within UWN are associated to all their corresponding meanings and therefore is possible to select words that appear associated to few different meanings, for training the CCA projections. These words with less different meanings are perhaps more likely to be used in the same contexts, across the different languages, are thus better candidates for establishing cross-language correlations. Table 3.4 presents the obtained results when considering 300-dimensional projections for the words that have less than two, or less than three different meanings within UWN, showing that higher correlations can be achieved in some few cases, although not consistently for all languages and/or types of forecasting models. The values from Table 3.4 that are shown in bold correspond to those cases where better results could be achieved by the smaller sets of seed translations, in comparison to those that are reported on Table 3.2. It should nonetheless be noted that this final set of experiments used the same values for the k, α and γ parameters that were considered for the experiments reported on Table 3.2. Slightly better results can perhaps be achieved by tuning these parameters (e.g., through grid-search).

3.5 Discussion

Human ratings for affective variables associated to lexical units, such as valence, arousal or dominance, are a fundamental resource for many fields of research, but they are also notoriously

difficult to collect. Although it is nowadays possible to efficiently obtain measurements for tens of thousands of words, e.g., by using crowd-sourcing platforms (Warriner *et al.*, 2013a), automated approaches for establishing such norms can still offer an interesting alternative.

In this chapter, was presented a novel technique for estimating lexical norms automatically, based on training predictive models from existing norms, and leveraging word embeddings obtained through modern approaches based on neural networks. Moreover, was also proposed a technique based on canonical correlations analysis, for leveraging English data with the purpose of estimating lexical norms for other languages (i.e., Spanish, Portuguese, Italian and German). Experimental results attest to the effectiveness of the proposed approach. This method surpasses the current state-of-the-art when predicting emotional ratings for English words and, in cross-language experiments, was also measured a significant correlation with human-ratings.

Although our experiments have shown promising results with the usage of Canonical Correlations Analysis (CCA) for leveraging English data with the purpose of estimating lexical norms for other languages, we have that this particular technique is only able to reveal linear relationships in the data. Recent studies have introduced efficient methods, essentially corresponding to extensions of the linear method known as CCA, to learn complex nonlinear transformations of two views of data, such that the resulting representations are highly linearly correlated (Andrew *et al.*, 2013; Lopez-Paz *et al.*, 2014). These approaches have been shown to find representations with significantly higher correlation than those learned by CCA and, for future work, it would be interesting to experiment with similar non-linear approaches, as a way of building improved representations for the task of making cross-language predictions of lexical norms.

	ANEW		
	Valence	**Arousal**	**Dominance**
Spanish	0.92	0.75	0.72
Portuguese	0.91	0.58	0.67
Italian	0.92	0.63	0.75
German	0.89	0.63	0.60

	Warriner et al.		
	Valence	**Arousal**	**Dominance**
Spanish	0.92	0.69	0.83
Portuguese	0.91	0.57	0.67
Italian	0.92	0.62	0.76
German	0.91	0.66	0.70

Table 3.3: Correlations between human norms for English words and human norms in the four different adaptations of the ANEW lexicon. The corresponding p-values were always lower than 0.001.

	Seed	k-NN			Random Forests			Kernel Ridge		
	Words	Val	Aro	Dom	Val	Aro	Dom	Val	Aro	Dom
<2 meanings										
Spanish	8893	0.583	0.396	0.584	0.619	0.388	**0.601**	0.666	0.438	0.624
Portuguese	6619	**0.572**	0.396	**0.455**	0.572	0.434	0.452	0.656	0.384	**0.510**
Italian	8907	0.573	0.410	0.460	0.573	0.421	0.500	0.623	**0.457**	0.504
German	2640	0.374	0.210	0.284	0.412	0.373	0.249	0.443	0.283	0.295

	Seed	k-NN			Random Forests			Kernel Ridge		
	Words	Val	Aro	Dom	Val	Aro	Dom	Val	Aro	Dom
<3 meanings										
Spanish	12240	0.592	0.418	0.564	0.630	0.400	**0.599**	0.674	**0.460**	0.624
Portuguese	8961	**0.569**	0.378	0.437	0.582	**0.450**	0.465	0.649	0.367	0.503
Italian	12538	0.563	**0.430**	0.468	0.592	0.422	0.505	**0.644**	0.442	**0.529**
German	4357	0.465	0.292	0.367	0.461	0.429	0.318	0.530	0.380	0.356

Table 3.4: Obtained results, in terms of Pearson's correlation coefficient, when using different versions of the seed lexicon containing word translations for projecting the word embeddings according to CCA. The corresponding p-values were always lower than 0.001.

	Valence	**Arousal**	**Dominance**
Spanish	0.700	0.459	0.653
Portuguese	0.665	0.420	0.443
Italian	0.679	0.494	0.569
German	0.617	0.619	0.527

Table 3.5: Obtained results, in terms of Pearson's correlation coefficient, when using monolingual data through a leave-one-out cross validation methodology. The corresponding p-values were always below 0.001.

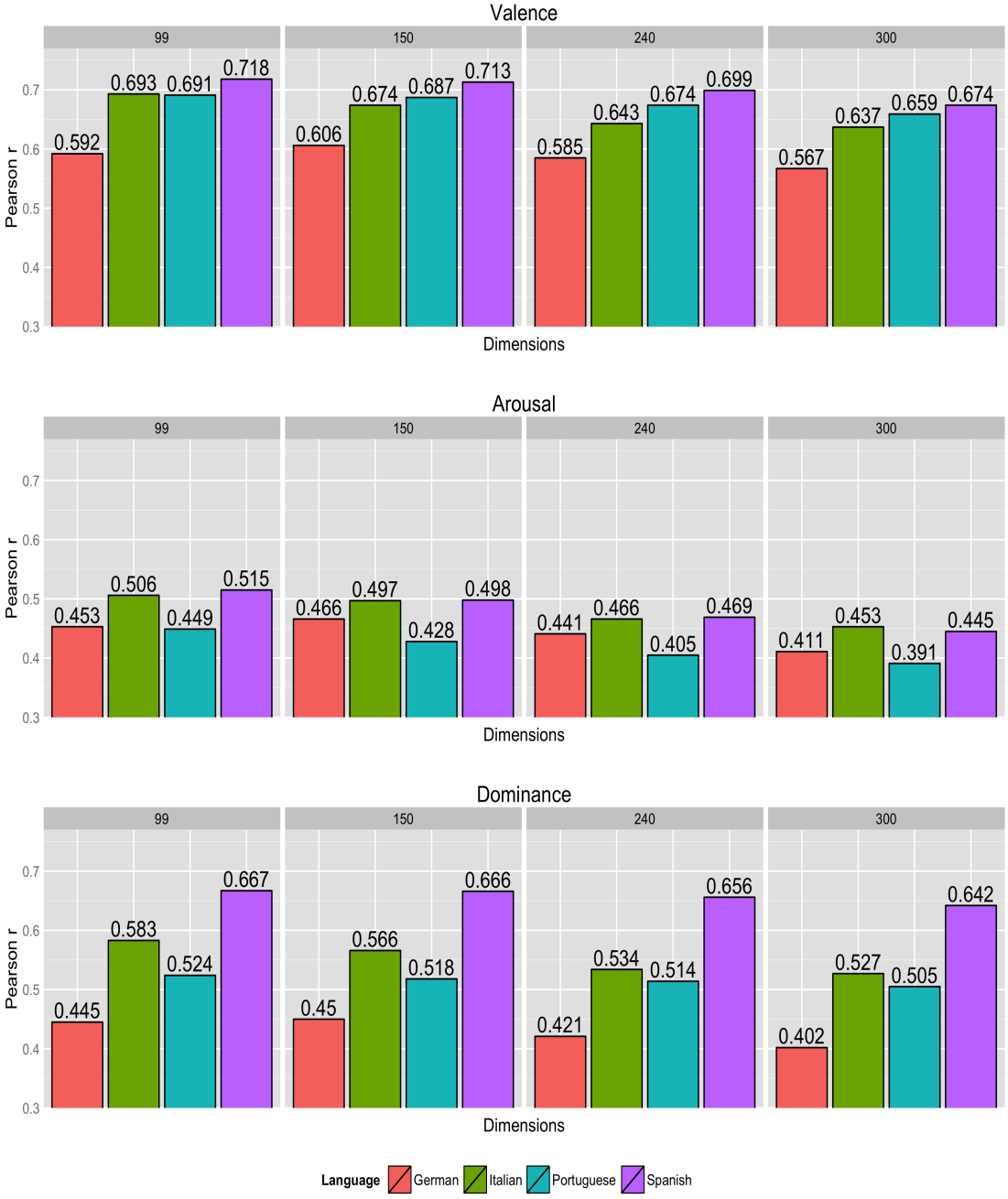

Figure 3.5: Obtained results in terms of Pearson's correlation coefficient, for the different languages when taking only the top n most correlated dimensions that are produced by the CCA projections.

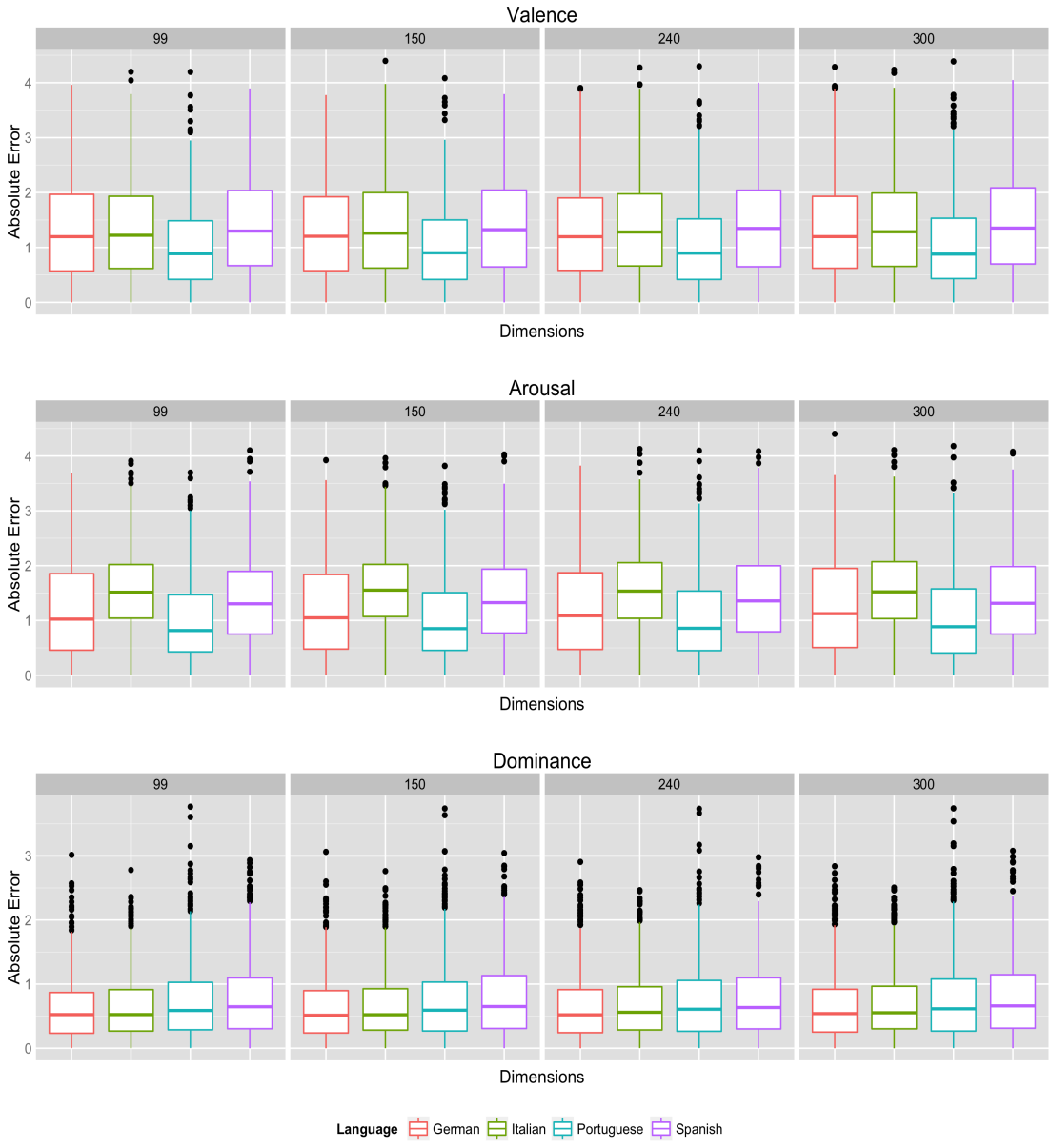

Figure 3.6: Obtained results when taking only the top n most correlated dimensions that are produced by CCA, in terms of the distributions for the absolute errors that were measured.

Chapter 4

Predicting Affective Norms for Short Texts

This chapter reports on an empirical evaluation of an automated method for predicting emotional responses to text items, according to dimensions of valence, arousal and dominance. The considered method is based on embeddings for words and larger pieces of text, obtained through unsupervised approaches based on neural networks. Leveraging these embeddings as features, together with existing norms from Bradley & Lang (1999) for a large set of English words, were trained models for estimating emotional ratings for both previously unseen words or documents. Experimental results attest to the effectiveness of this approach, showing very high correlations towards human ratings, for both words and documents.

4.1 Introduction

Despite much sentiment analysis research, few previous studies are directly related to predicting responses to text items in a $[1 - 9]$ scale according to dimensions of valence (i.e., pleasantness of the stimulus), arousal (i.e., intensity of feeling being evoked) and dominance (i.e., the degree to which the stimulus makes the reader feel *in control*).

Paltoglou *et al.* (2013) have previously reported on a study in which subjects were asked to rate the emotional impact of 20 forum discussion posts, according to valence and arousal. These documents were later analyzed through the ANEW lexicon, estimating the emotional content of each document through the weighted geometric mean of the ANEW tokens found it the text. The authors report on correlations with the human assessments of 0.89 for valence and of 0.42 for

arousal.

Murphy (2014) also experimented with a simple procedure based on assigning documents to the average ANEW scores of their words, validating their results on data from the Affective Norms for English Texts (ANET) dataset (Bradley & Lang, 2007). The authors report on correlations of 0.45 for valence, 0.31 for arousal, and 0.26 for dominance.

The methods presented in this subsection differ significantly from the aforementioned approaches, by going beyond rule-based methods that use lexicons for text analysis, and by leveraging unsupervised embeddings for words and larger pieces of text (Le & Mikolov, 2014). The availability of large sets of training documents, rated according to valence, arousal and dominance, is a serious limitation to the development of text mining methods for estimating emotional ratings for texts. The proposed method addresses this particular limitation, given that is instead proposed to leverage the existing ratings for words, as training data for building predictive models (e.g., the embeddings for words such as *candy* and *love* should probably be similar to an embedding for the sentence *everyone loves candy*, and thus predictive models trained with basis on ratings for the words can perhaps be used to infer ratings for longer pieces of text). Moreover, given that emotional ratings similar to those of the ANEW dataset are readily available for many different languages (Montefinese *et al.*, 2014; Soares *et al.*, 2012), the approach proposed here is also readily applicable to non-English text, allowing us to effectively associate real-valued emotional ratings to textual contents in general.

4.2 Using Paragraph Embeddings for Predicting Sentence Ratings

This experiment is based on the idea that emotional norms for English words, such as those made available by Warriner *et al.* (2013a), can be used as training data for predictive models capable of estimating valence, arousal and dominance for previously unseen (sequences of) words. The aim is using embeddings, obtained through unsupervised approaches based on neural networks, as features within different types of regression approaches.

Following the success of word embedding techniques such as those from word2vec, researchers have tried to extend these models to go beyond the word level, specifically aiming to achieve phrase-level or sentence-level representations. A simple approach involves using a weighted average of the embeddings for all the words in the document, losing the word order in the same way as the standard bag-of-words models do. However, authors like Le & Mikolov (2014) have proposed more sophisticated approaches, consisting of unsupervised frameworks, similar to those

from word2vec, that learn representations for variable-length pieces of text. In the *paragraph vector* approach from Le & Mikolov (2014), the representations are trained for predicting words in a document (i.e., the authors concatenate the *paragraph vector* with several word vectors from a document, and predict the following word in the given context). While *paragraph vectors* are unique among documents in a given corpus, the word vectors are shared and can be trained separately on larger corpora, afterwards keeping them fixed while training the paragraph vectors by stochastic gradient descent and back-propagation, until convergence.

The aforementioned *paragraph vector* approach is very similar to the skip-ngram method from word2vec. Each *paragraph* can be thought of as another word, acting as a memory that remembers what is missing from the current context (i.e., the topic of the document). Notice also that in the *paragraph vector* approach, both the words and the documents end up being embedded in a common vector space. We can therefore train predictive models leveraging word embeddings in the same vector space as the documents, in order to estimate properties for entire documents.

Specifically, predictive models were trained, using the same algorithms from the previous chapter (i.e., k-NN, Kernel Ridge and Random Forests) and the lexical norms made available by Warriner *et al.* (2013a), for the prediction of emotional ratings for entire documents. *Paragraph vectors* were trained for the document collections used in our experiments keeping the word embeddings fixed according to those from word2vec's website. The implementation from the GenSim package[1] was used for training the 300-dimensional *paragraph vectors*, using the default parameters from this library. Prior to model training, the *paragraph vectors* were initialized according to a weighted average of the embeddings for the words occurring in the documents, using TF-IDF scores as the word weights.

4.3 Results

It was attempted to see if the norms for 13,915 English words previously made available by Warriner *et al.* (2013a), could be used to estimate valence, arousal and dominance for entire documents, through unsupervised embeddings generated through the *paragraph vectors* approach (Le & Mikolov, 2014).

Thus, were trained predictive models using the entire dataset of 13,915 English words from Warriner *et al.* (2013a), and evaluated the results through the Affective Norms for English Text (ANET) dataset, which provides a set of normative emotional ratings for a total of 120 brief texts in the English language (Bradley & Lang, 2007).

[1] `http://radimrehurek.com/gensim/`

	ANET					
	Valence		Arousal		Dominance	
	Pearson	MAE	Pearson	MAE	Pearson	MAE
Random Forest	0.594	2.201	0.419	2.476	0.583	1.540
k**-NN**	0.647	2.090	0.569	2.351	0.627	1.499
Kernel Ridge	0.732	1.809	0.649	2.067	0.680	1.392

	EmoTales					
	Valence		Arousal		Dominance	
	Pearson	MAE	Pearson	MAE	Pearson	MAE
Random Forest	0.275	0.999	0.134	1.472	*0.045*	0.876
k**-NN**	0.330	1.008	0.177	1.424	0.100	0.872
Kernel Ridge	0.347	1.087	0.217	1.418	0.112	0.880

Table 4.6: Results when predicting ratings for the texts in the ANET (Bradley & Lang, 2007) and EmoTales (Francisco *et al.*, 2012) datasets. The p-values for Pearson's correlations were always lower than 0.0001, except for the case shown in italics.

Table 4.6 presents the obtained results, in terms of Pearson's ρ and in terms of the Mean Absolute Error (MAE). The parameters associated to the k nearest neighbor (i.e., $k = 100$) and kernel ridge regression (i.e., $\alpha = 1$ and $\gamma = 1$) approaches were tuned through a simple grid-search, so as to optimize the average scores in all three emotional dimensions. The results, although inferior to those obtained for predicting the ratings of individual words, attest to the high effectiveness of this method. In fact, the results obtained by this method are significantly superior to those from previous comparable studies (Murphy, 2014). It should nonetheless be noted that the correlations between the ratings for male and female subjects, in the original ANET study, are of approximately 0.98 for valence, and of 0.96 for both arousal and dominance. This suggests that there is still room for significant improvements.

In a second set of experiments, was used a previously available dataset, consisting of 20 forum posts rated according to valence and arousal (Paltoglou *et al.*, 2013). The emotional norms associated to the documents in this dataset were first converted into the interval $[0 - 9]$. Predictive models, were then trained using the entire dataset of 13,915 English words from (Warriner *et al.*, 2013a), finally evaluating the results in terms of Pearson's ρ and in terms of the MAE. The parameters associated to the k nearest neighbor (i.e., $k = 18$) and kernel ridge regression (i.e., $\alpha = 10$ and $\gamma = 1$) approaches were once again tuned through grid-search, so as to optimize the average scores in both dimensions.

Table 4.7 presents the obtained results, where we can again see that the proposed methodology is able to rate texts with a reasonably high accuracy. The method based on embeddings outperforms the results reported by Paltoglou *et al.* (2013) for the dimension of arousal, although not in terms of valence.

	Forum Posts			
	Valence		Arousal	
	Pearson	MAE	Pearson	MAE
Rnd. Forest	0.753	1.338	0.654	0.667
k-NN	0.627	1.435	*0.193*	0.851
Kernel Ridge	0.791	1.720	0.785	0.377

Table 4.7: Results obtained when predicting ratings for texts from forum posts. The p-values for the Pearson correlations were lower than 0.001, except for the case shown in italics.

Finally, in a third set of experiments, was used the *EmoTales* dataset composed of a total of 1168 sentences belonging to 18 folk tales (e.g., *Cinderella*, *Rapunzel*, etc) and rated according to valence, arousal and dominance by 36 human annotators (Francisco *et al.*, 2012). Then, as done with the two previous set of experiments, regression models were trained using the dataset from Warriner *et al.* (2013a) and the results were evaluated in terms of Pearson's ρ and in terms of the MAE. The parameters associated to the k nearest neighbor (i.e., $k = 100$) and kernel ridge regression (i.e., $\alpha = 1$ and $\gamma = 0.01$) models in this experiment were also optimized through a simple grid-search approach, in order to optimize the average scores in terms of predicting valence, arousal and dominance.

Table 4.6 presents the obtained results when predicting the ratings of the sentences that compose the *EmoTales* dataset (Francisco *et al.*, 2012). The results were significantly lower than the ones obtained on the previous experiments presented in this article. These inferior results, may be assigned to the fact that in this dataset the meaning of each individual sentence, and thereafter its emotional rating is not independent from the other sentences that compose the tale. For example, in the *The Wicked Prince* tale, where the prince is the evil character in the tale, the sentence *crept into the prince's ear and stung him* was rated with a valence of 6.67, since the annotators considered this action against the prince to be a positive one, besides the sentence being composed of negative terms. This specific characteristic of this dataset makes it very difficult for our method to correctly predict the ratings of the sentences that compose this dataset, since these ratings are highly dependent of context information.

4.4 Discussion

In this chapter, were empirically evaluated automated approaches for predicting emotional responses to textual contents, according to emotional dimensions of valence, arousal and dominance, as defined within the well-known ANEW study. When inferring ratings for entire textual documents, were achieved correlations of 0.732 in terms of valence, 0.649 in terms of arousal,

and 0.680 in terms of dominance, in tests with documents from the ANET corpus. Moreover, when inferring emotional scores for forum posts, were achieved correlations of 0.79 and of 0.785 in terms of valence and arousal, respectively.

Despite the interesting results, there are also several possible paths for improvement. For future work, would be positive to extend our experiments to other languages besides English, either by leveraging existing emotional norms for words in other languages, or by leveraging cross-language word embeddings in order to transfer norms for large sets of English words into other languages (Faruqui & Dyer, 2014). Finally, besides word2vec embeddings, would also be worth to experiment with alternative word embedding procedures (Liu *et al.*, 2015; Pennington *et al.*, 2014).

Chapter 5

Affect and Emotions over Twitter Messages

This chapter describes two different applications that leverages on the prediction of emotional scores (i.e., in terms of valence, arousal and dominance) for tweets. The method used to predict the aforementioned scores is described, in detail, in Chapter 4. The first application, presented in Section 5.2, tries to predict the well-being of populations across continental USA (excluding Alaska), making use of geo-referenced tweets. The second and last application, described in Section 5.3, consists in seeing if there is correlation between the emotional scores of a given tweet and the metrics associated to the diffusion of that same tweet across the social network.

5.1 Introduction

The convergence point of the applications described in Section 5.2 and in Section 5.3 lies in the usage of a dataset of Tweets, and in the prediction of emotional scores in terms of Valence, Arousal and Dominance to the textual content of these same tweets.

An already existent dataset of 325.333.833 tweets posted by 19.558.917 different users, already used by Loff *et al.* (2015), was leveraged on these applications. The tweets in this dataset were collected from the Twitter streaming API service, during the year of 2012. Some characterization statistics are presented in Table 5.8, such as: (i) the percentage of tweets which have an associated pair of geographic coordinates, (ii) the percentage of tweets which are retweets, and (iii) the percentage of tweets which are replies.

	All Tweets			
	Tweets	**Retweets**	**Replies**	**Geo-referenced**
Count	325.333.833	129.800.772	105.151.446	1.734.978
%	100	39,90	32,32	0,53
	Tweets in English			
	Tweets	**Retweets**	**Replies**	**Geo-referenced**
Count	174.331.835	79.278.776	48.114.205	990.402
%	100	45,48	27,60	0,57
	Tweets in English and from continental USA (excluding Alaska)			
	Tweets	**Retweets**	**Replies**	**Geo-referenced**
Count	449.647	1	219.715	449.647
%	100	-	48,86	100

Table 5.8: Statistical characterization of the Twitter datasets.

One of the divergence points between the two applications, is related to set of tweets used in the experiments. Both applications use only the subset of tweets which are written in English (i.e., which have the tag *en* associated to the language of the author). But while the application described in Section 5.3 makes usage of all the tweets in the aforementioned dataset, the application presented in Section 5.2 uses only those messages which already have an associated pair of coordinates (e.g., messages posted from mobile phones that have a GPS receiver), and whose coordinates lie within the USA Continental states (excluding Alaska).

Nevertheless, the methodology followed to predict the emotional scores of these tweets, which will later be used for different purposes, was the same. This methodology is highly supported by the results presented in Chapter 4, where experiments on the prediction of Valence, Arousal and Dominance for textual contents, based on their embeddings, were made.

In order to gather all the pre-requisites to predict the emotional scores of the tweets in the aforementioned sets, it began by training a word2vec model with all the words from all the tweets written in English. All numbers and user mentions in the tweets were replaced by a symbol (e.g., *NUMBER*). The main goal was to leverage on a word2vec model which contained all the specific terms and words specific to tweets (e.g., emoticons, hashtags and keywords), since other publicly available word2vec were trained using other types of texts. For example, the word2vec model from Google, was trained using a news dataset. The implementation used to train the word2vec model was the one from the GenSim package, and the size of the embeddings was set to a dimensionality of 300. A window size of 5 was used while training. So that the robustness of

	ANEW					
	Valence		Arousal		Dominance	
	Pearson	MAE	Pearson	MAE	Pearson	MAE
Random Forest	0.647	1.407	0.389	0.827	0.560	0.709
k-NN	0.680	1.551	0.372	0.831	0.522	0.768
Kernel Ridge	0.725	1.151	0.457	0.765	0.605	0.652

	Warriner et al.					
	Valence		Arousal		Dominance	
	Pearson	MAE	Pearson	MAE	Pearson	MAE
Random Forest	0.526	0.901	0.235	0.701	0.429	0.692
k-NN	0.575	0.894	0.358	0.676	0.468	0.695
Kernel Ridge	0.629	0.787	0.395	0.661	0.516	0.648

Table 5.9: Results when predicting ratings for the words in the ANEW and in the emotional lexicon from Warriner *et al.* (2013a), following a 10-folds cross validation approach. The p-values for Pearson's correlations were always lower than 0.0001.

these embeddings could be evaluated, i.e., to measure if these embeddings captured somehow the emotional semantics of the words they refer to, was performed a 10 folds cross-validation over two different emotional lexicons (i.e., ANEW and the one from Warriner *et al.* (2013a)), in which the embeddings of a given word in these lexicons was used as feature in a regression model that predicted the emotional scores as stated in the aforementioned lexicons. Random Forests, k-Nearest Neighbours (k-NN) and Kernel Ridge were used as regression models. These models have some parameters liable of being tuned: (1) In the case of Random Forests, the number of forests was set to the dimension of the embeddings. (2) In the case of k-NN the number of neighbours k may be adjusted. (3) And finally in the Kernel Ridge regression model, α and γ may also be subject to optimization. These last three parameters were subject to a simple grid search optimization, in which was tried to optimize the quality of the models when considering the Pearson correlation coefficient and the MAE between the predicted scores and the scores from the aforementioned lexicons. Table 5.9 presents the results, described in terms of Pearson correlation coefficient and MAE, obtained when training and evaluating the models through a 10-folds cross validation methodology. These results although being inferior to those obtained when predicting the ratings of unseen words (see Table 3.1) when leveraging on the embeddings provided by Google, are still robust enough to sustain further use of these embeddings for other tasks.

Then, paragraph vectors were trained for the set of tweets considered by the two applications. The implementation used to train these paragraph vectors was also the one from GenSim package, using the default parameters and setting the size of the embedding to 300. Before training the model, the embeddings of words were initialized with the values from the word2vec model

described in the last paragraph, and the vectors associated to each one of the tweets were initialized according to a weighted average of the embeddings of the words that composed the tweet. TF-IDF scores were used as word weights.

Having the paragraph vectors associated to the aforementioned set of tweets, and following the methodology presented in Chapter 4 where predictive models of the emotional scores for unseen documents were built, was trained a Kernel Ridge regression model using the lexicon from Warriner *et al.* (2013a) and the embeddings from the previously trained GenSim doc2vec model. This regression model, based on the embedding of a given word/tweet is able to predict its emotional score. While training the Kernel Ridge regression model, the parameters subject to tuning (i.e., α and γ) were set to the ones found to be optimal in the aforementioned 10-folds cross-validation experiment with the lexicon from Warriner *et al.* (2013a). This way, the emotional score of each tweet considered for this experiment, was predicted, and able to be used in the experiments of Section 5.2 and Section 5.3.

As a simple way of visualizing the predictions made for the tweets, and to see whether these predictions are in line with the boomerang visual pattern formed by the correlation between Valence in Arousal, in emotional lexicons such as the ANEW or the one from Warriner *et al.* (2013a), were plotted side-by-side, in Figure 5.7, the correlation between Valence in Arousal, in the ANEW emotional lexicon, in the lexicon from Warriner *et al.* (2013a), and in the predictions made for the datasets used in the two applications.

5.2 Predicting Well-Being with Twitter

This section proposes on leveraging geo-referenced social media data extracted from Twitter within the year of 2012, together with features based on embeddings of tweets and on emotional scores predicted to these same tweets, to estimate well-being of populations from specific geographic points, across continental USA. Namely, as an extension to the work of Loff *et al.* (2015), was attempted to produce a regression model, that based on the aforementioned features, estimates a composite well-being index (i.e., Gallup-Healthways composite index) built leveraging on telephone-based interviews. These interviews were composed of questions about themes that ranged from health behaviours, physical and emotional health, work environment to themes such as financial security and access to basic needs (i.e., shelter, water, food, health-care). Well-being of populations translates into how these populations evaluate their lives (i.e., life satisfaction) and their general emotional state (i.e., do they general feel positive or negative emotions?). This method addresses the disadvantage that arises from surveying as mean of assessing happiness

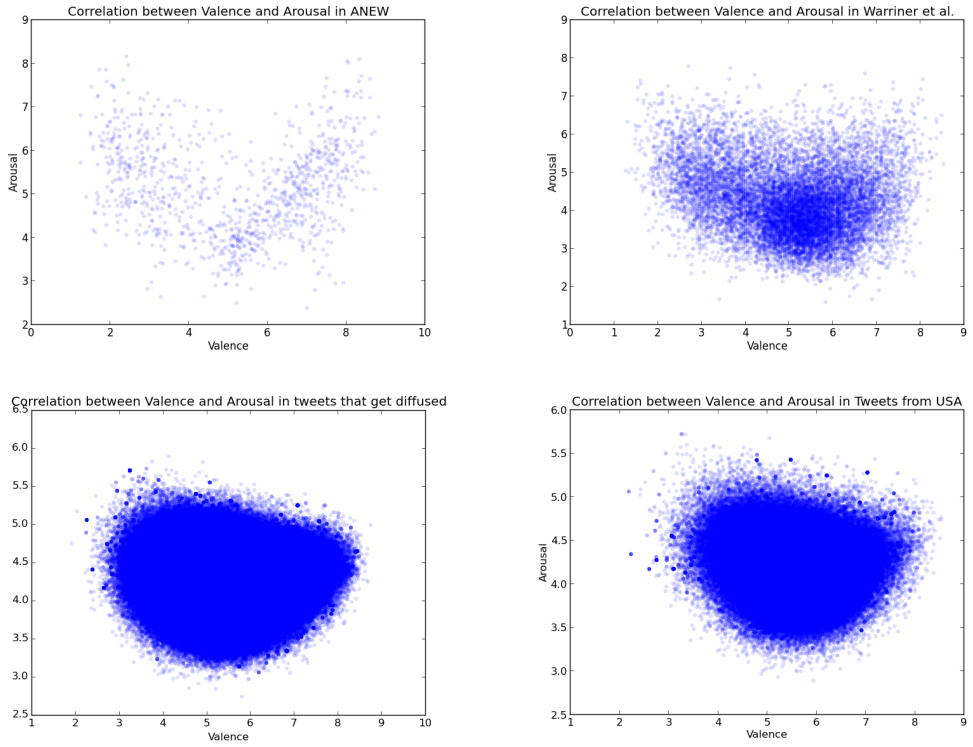

Figure 5.7: Correlation between valence and arousal in the ANEW lexicon (top left), in the lexicon from Warriner *et al.* (2013a) (top right), in the dataset of tweets used in the Gallup-Healthways well-being score prediction experiment (bottom right), and in the dataset of tweets used to reconstruct information diffusion processes (bottom left).

levels of populations, since it is economically expensive and time consuming, by being able to extract these same information relying solely on data extracted from social networks.

Being in line with the approach followed by Loff *et al.* (2015), and using the same dataset composed of tweets extracted from Twitter during the year of 2012, were only used the tweets whose geographic coordinates were already present in its metadata, and was decided not to infer the location of a tweet by using heuristics such as the usage of location field of the user that produced a given message. This decision allowed this study be similar to (Loff *et al.*, 2015), in the sense that it uses the same type of linear regression model (i.e., Elastic Net) and leverages on the same dataset, the major difference relies on the set of features used to train the model, i.e., while Loff *et al.* (2015) relied on features that consisted in the word-counting of some emotional keywords, such as the ones present in the ANEW emotional lexicon, this method leverages on features such the phrase embeddings of the tweets produce by a given state, as well as, the emotional scores of these same tweets predicted using the method presented in the previous section.

The general approach consisted in following a leave-one-out cross validation methodology to

train and evaluate a linear regression model that would predict for each state (in continental USA, excluding Alaska) the well-being overall score in the individual state-level reports of the Gallup-Healthways well-being composite index relative to the year of 2012. In 2012, the national average for the well-being overall score was of 66.5 in a range of $[0, 100]$. The highest score was achieved by the state of Colorado (69.4) and the lowest was achieved by the state of West Virginia (61.3).

Leveraging both on the paragraph vectors, and on the scores estimated for each one of the considered tweets it was able to produce features that would later be used in the well-being predictive models. Firstly, the tweets issued from each one of the continental USA states (excluding Alaska), were aggregated, then for each one of these states, the following features were calculated:

1. The *average*, and the *median* of each one of the elements from the 300-dimensional embeddings, among all the tweets belonging to that given state;
2. The *maximum*, the *minimum*, the *average* and the *mode* of the previously predicted Valence, Arousal and Dominance scores for all the tweets associated to that given state;
3. The number of tweets of associated to a given state, N;
4. $\frac{N}{T}$, where T is the total number of tweets considered in the experiment.

Then, 3 different regression models, were built, making use of the aforementioned features for each one of the states in the continental USA (excluding Alaska), and using the standard scores for each state in the Gallup-Healthways composite well-being index. A linear regression model with Elastic Net regularization was learned, with the aim of predicting well-being scores over the aforementioned states. The 3 regression models differ in the set of features they leverage on, i.e., some of the models rely only on a subset of the aforementioned set of features. This allowed me to assess which kind of features (i.e., those related to embeddings, those related to scores, or a combination of both) contributed most to the prediction of well-being. To easily distinguish the models, was associated a letter to each one of them: (1) *Model A* leveraged features 1,2,3 and 4; (2) *Model B* leveraged features 2, 3 and 4; And finally *Model C* leveraged features 1, 3 and 4. *Model A* tries to combine both scores and embeddings, while *Model B* focus on using only scores, and *Model C* focus on using solely embeddings.

A linear regression model with Elastic Net regularization was chosen primarily for two reasons: (1) For maintaining the maximum resemblance to the methodology followed by Loff *et al.* (2015), since the authors also used a linear regression model with Elastic Net regularization. Thus, making the following results highly comparable with the ones from Loff *et al.* (2015). And (2) due to the fact that the number of considered features per state is of 614 ($300 \times 2 + 3 \times 4 + 2$)

	Pearson ρ	**Kendall τ**	**MAE**	**RMSE**
Model *A*	0.759	0.575	0.866	1.193
Model *B*	0.434	0.336	1.217	1.581
Model *C*	**0.772**	**0.590**	**0.839**	**1.163**

Table 5.10: Results obtained by comparing the predictions from Models *A, B* and *C* to the ground-truth scores from Gallup-Healthways well-being index.

in case of *Model A* and of 602 in case of *Model C*. Elastic Net regularization is suitable for these situations (i.e., high number of features) since this regularization method selects the most predictive features among the others. (For a deeper understanding of this regularization method please see Section 2.2.1 where the work of Loff *et al.* (2015) is described in more detail.) For training the model parameters, was used the implementation also used by Loff *et al.* (2015), i.e., the one from glmnet package for the R system for statistical computing.

So that the models could be evaluated in terms of the prediction of the well-being scores according to the Gallup-Healthways well-being index, a leave-one-out cross validation methodology was performed, in which all the states except one are used to train the regression models, and the score associated to remaining one is predicted. This method is repeated for each one of the considered USA states, and at the end, when predictions for each one of these states were available, the results were evaluated by comparing the predictions to the ground-truth scores from Gallup-Healthways well-being index. The predictions and the ground-truth are then compared leveraging on the following metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Pearson correlation coefficient ρ and Kendall correlation coefficient τ. These correlation metrics are useful to see if, the states, when ranked according to the predictions or according to the ground-truth, appear in similar positions. The obtained results, in this leave-one-out cross validation experiment, are presented in Table 5.10. Please note, that a baseline approach consisting in assigning the average score of all states to every state would achieve a MAE of 1.40 and RMSE of 1.73. The obtained results with *Model C* (i.e., a MAE of 0.839, a RMSE of 1.163, a ρ of 0.772, and finally a τ of 0.590.), are consistently superior (except in the case of RMSE) than those reported by Loff *et al.* (2015), i.e, a MAE of 0.92, a RMSE of 1.22, a ρ of 0.74, and finally a τ of 0.58. These results suggest that the method of inferring well-being of populations based on embeddings of geo-referenced tweets is robust.

In Table 5.11, are presented for each considered state, the ground-truth well-being score, the well-being score predicted by Model *A, B* and *C*, the number of geo-referenced Tweets issued from that given state, and finally the predicted average Valence, Arousal and Dominance scores.

Finally, in order to present a visualization of the results, Figure 5.8 presents, side by side, four choropleth maps of continental USA, produced using the R system for statistical computing.

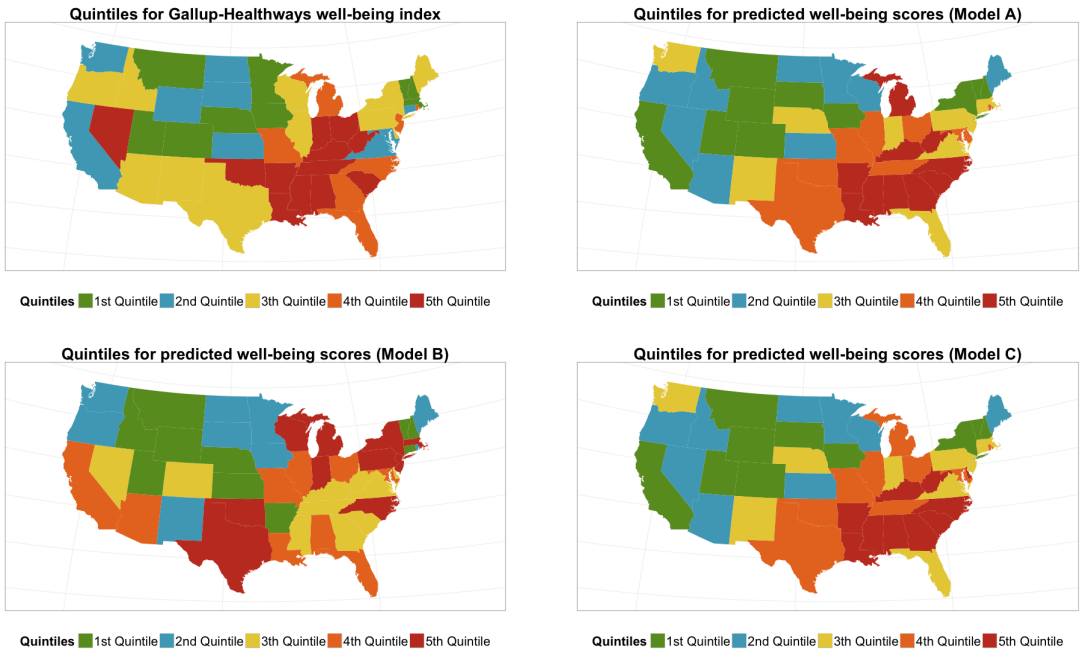

Figure 5.8: Per-state well-being in continental U.S. The top left map was built using the reports by Gallup-Healthways and the others were built leveraging on the regression models predictions.

These maps, present the results in terms of quintiles, i.e., the domain of well-being scores was divided in 5 parts of equal-size. Each one of the colors represent a given quintile. One of the maps was built using the ground-truth data, while the others were built using the predictions made by the regression models.

5.3 Correlating Information Propagation with Affective Ratings

Quantifying the nature, intensity, and geographic distribution of emotional states, at the level of populations, is important to better construct public policy, and also from a scientific perspective to more fully understand economic and social phenomena. Emotions play a major role in human interactions, and therefore the hypothesis that emotions may also interfere with the way users interact with each other, in the context of online social networks, is relevant. Questions such as: (1) *do negative messages in social networks spread faster and further than positive ones*, or (2) *does the emotional valence of a message affect the countries where a message is most spread*, or even (3) *is the gender of the users who spread a given message related to the emotional valence of a message* are some of the questions that are intended to answer throughout this

section.

In order to correlate the way content spreads in a social network with the emotional scores expressed in that content, the information propagation process originated by the initial messages will be reconstructed. After that, some properties (e.g., depth, width, or geographical coverage) of these processes may be extracted and analyzed. Finally, correlations between these properties and the emotional valence will be tested.

The reconstruction of the diffusion processes imposes particular challenges, since the relations of causality between retweeted messages, needed to establish how the messages spread, are not explicitly expressed in a dataset of messages extracted from Twitter. In Twitter, a retweeted message does not contain in its metadata the id of the message from where it has been retweeted from. This leads to a set of possible messages from which the author may have retweeted. Therefore, to reconstruct the diffusion process, a set of heuristics must be employed, corresponding to a propagation model. This set of heuristics ranges from the geographical proximity between the author and other users, to their historical interactions.

Seeing if the emotional scores of messages (i.e., in terms of valence, arousal, and dominance) affects the way these same messages spread (e.g., how far they get) in social networks like Twitter may be useful, for instance, to marketing purposes, since marketeers may adapt the emotional content of a message to better fulfill a given campaign goal.

Regarding the methodology followed to pursue the aforementioned goal, it began by gathering the predictions of the emotional scores of the tweets (from the aforementioned dataset) that happened to generate retweets. Then, leveraging on a algorithm which tries to reconstruct the information propagation process (i.e., the chain of retweets) generated by a given tweet, some fluxes of retweets are reconstructed, and on top of these fluxes, diffusion metrics (e.g., depth, width, geographic range, etc) are calculated. Finally, having the emotional scores associated to a tweet that happened to generate a information propagation process, and the metrics associated to this same process, for the set of all reconstructed processes, the relations between each one of the considered emotional dimensions (i.e., valence, arousal, and dominance) and each one of the considered diffusion metrics, were plotted.

Concerning the extraction of the information propagation process originated from a given message posted in Twitter, was taken into account the fact that, in Twitter, these processes are not explicitly expressed. In Twitter, a retweeted message does not contain in its metadata the id of the message from which it was retweeted, but solely the id of the message that is in the origin of all the retweets. Thus, is required to reconstruct the chain of retweets making use of the combination of some heuristics corresponding to different propagation models. If there is a message

Algorithm 1: Similarity between two tweets for chain of retweets reconstruction

Input: pair of messages $tweet1$ and $tweet2$
Output: decimal number representing their similarity
$timestamp1 \leftarrow tweet1.timestamp$;
$timestamp2 \leftarrow tweet2.timestamp$;
if $timestamp1 \leq timestamp2$ **then**
  return 0;
end
$user1 \leftarrow tweet1.user$;
$user2 \leftarrow tweet2.user$;
$mentions \leftarrow user1.mentioned(user2) \vee user2.mentioned(user1)$;
$retweets \leftarrow user1.retweetedFrom(user2) \vee user2.retweetedFrom(user1)$;
if $\neg(mentions \vee retweets)$ **then**
  return 0;
end
$timestampSimilarity \leftarrow 1/(1 + timestamp1 - timestamp2)$;
$intersection \leftarrow user1.hashtags \cap user2.hashtags$;
$union \leftarrow user1.hashtags \cup user2.hashtags$;
$hashtagSimilarity \leftarrow intersection.length/union.length$;
$locationSimilarity \leftarrow 1/(1 + haversineDistance(tweet1.coords, tweet2.coords))$;
return $\alpha * hashtagSimilarity + \beta * timestampSimilarity + \gamma * locationSimilarity$;

m that is a retweet and was published by a user u who in the past has interacted with the set of users U, the message r from which m has been retweeted must be in the set of messages published from the users in U. To select the user $u' \in U$ that is the author of r, is considered criteria such as the geographical proximity to u, the level of previous interactions with u and the amount of shared hashtags with u. The timestamp of m must also be more recent than the timestamp of r, and similar to that of r. The pseudocode that materializes the calculation of the similarity between two tweets, for the purpose of reconstructing the diffusion process and taking into account the described heuristics, is expressed in Algorithm 1.

Having a graph G corresponding to the information propagation process originated by a given message, was developed an application that will support the calculation of properties (e.g., number of users, width, depth, duration and geographic coverage) of each one of the extracted information propagation processes. $G = (V, E)$ is a graph with a set of nodes V and a set of edges E. In G, the nodes correspond to the original message o and all the retweets of o. Each edge $e \in E$ connects a node i to a node j, where j represents the message from which the message corresponding to i was copied from. Depth corresponds to the number of edges that compose the longest path between o and any other node in G. Width is the maximum number of nodes whose size of the minimum path to o is equal. Duration consists in the difference between the timestamp of the last retweet in the process and the timestamp associated to the original tweet. Finally, geographic coverage is measured in two different ways: (1) The maximum distance in kilometers between any two nodes in G when applying the haversine distance approach to the geospatial

coordinates associated to each node. And (2) the area of the convex hull that encompasses the coordinates of all messages involved in the process.

After reconstructing each one of the diffusion processes, measuring all the proposed metrics, and finally associating every emotional score prediction of a tweet to the diffusion metrics of the diffusion process this same tweet has produced, was firstly plotted, the correlation for each considered emotional dimension (Valence in Fig. 5.9, Arousal in Fig. 5.10, and Dominance in Fig. 5.11) and for each considered diffusion metrics. By visually analyzing the plots, we cannot see any obvious correlation between any emotion dimension and any diffusion metric. It is also possible to infer that the emotional dimensions follow, as expected, a normal distribution, and that most diffusion metrics follow a exponential distribution, in which there is a high occurrences for low values, and very low occurrences for high values. To better confirm this last hypothesis was also presented, in Figure 5.12 for each diffusion metric, a histogram in a logarithmic scale. In these histograms, it is possible to confirm that all the metrics, except for the *longest distance between two tweets* metric, follow a exponential distribution. The fact that the *longest distance between two tweets* metric appears as a outlier, in terms of distribution, among the other metrics is possibly motivated by the low percentage of geo-referenced tweets in the dataset used in this experiment. This situation explains the histogram associated to the aforementioned metric, in which we can see high occurrences for diffusion processes with a distance of 0 and what seems to be a normal distribution for distances greater than 0.

As a last approach on trying to assess, if there was, even a slight, association between the considered diffusion metrics and the considered emotional dimensions, was plotted, in Figure 5.13, for each possible combination of diffusion metrics and emotional dimensions, the distribution of scores of the diffusion processes whose diffusion metric value was under the median/first quartile and the distribution of scores of the ones which were above the median/third quartile of the values for that given metric. Having the aforementioned plots, we can see that for all the metrics, except for *duration* and *maximum distance*, there is a tendency for the range of the distributions of scores above median/third quartile to be included in the range of the distribution of scores under median/first quartile (i.e., the maximum value of the first distribution is smaller than the maximum value of the second distribution and the minimum value of the first distribution is larger than the minimum value of the second distribution). In general, the are no differences between the medians of all the aforementioned distributions. From these observations we can infer a slight tendency for higher values for *depth, width, number of users*, and *area* to be associated to less variation of emotional scores in terms of valence, arousal, and dominance.

5.4 Discussion

Regarding the application described in Section 5.2, which consisted in predicting well-being scores for populations across continental USA, there are some limitations worth to point out, and possible paths of improvement for the future. The first one consists in the fact that embeddings by itself were found to be more predictive on the well-being of populations, than the emotional scores predicted on top of these embeddings. This suggest there are still space for improvement on what regards the prediction of emotional scores for tweets. Other limitation relies on the fact, that this study was based on tweets collected from the year of 2012. It would be interesting to see if these predictions maintain the same quality when reproduced for different years. Besides this limitation, the proposed approach is able to slightly outperform previous comparable studies

When looking into the application described in Section 5.3 where is intended to see if there is correlation between affective ratings of tweets and some diffusion metrics associated to the same tweets, we cannot draw strong conclusions. Apparently there is no strong relationship between these two factors in analysis. One of the main limitations in the methodology followed to study these relationships, was the fact that the method for reconstructing the path of diffusion of a given tweet across Twitter was not able to be tested. One way of testing it would be to apply this same methodology to datasets of messages from other comparable social networks where these diffusion paths are explicit. Unfortunately, access to that datasets was impossible.

State	Gallup-Healthways	Model A	Model B	Model C	Num Tweets	Avg Val	Avg Aro	Avg Dom
Alabama	64,2	65,5	66,2	65,4	7130	5,64	4,15	5,52
Arizona	67,1	67,8	66,2	67,3	6293	5,71	4,18	5,55
Arkansas	64,1	65,3	66,9	65,1	2906	5,63	4,16	5,51
California	67,4	67,3	66,2	67,5	39008	5,62	4,17	5,5
Colorado	69,7	67,4	66,5	67,4	2395	5,65	4,19	5,52
Connecticut	67,6	66,8	67	66,6	4272	5,62	4,17	5,49
Delaware	66,6	65,5	66,5	65,6	2163	5,6	4,17	5,48
Florida	65,8	66,5	66,2	66,6	26504	5,64	4,17	5,51
Georgia	66,1	65,3	66,2	65,3	31448	5,59	4,16	5,48
Idaho	67,1	67,4	68	67,1	334	5,79	4,23	5,6
Illinois	66,6	65,8	66,2	65,7	15340	5,61	4,17	5,49
Indiana	65,1	66,2	65,7	66,3	8594	5,69	4,18	5,54
Iowa	68,1	67,5	66,8	68	2922	5,49	4,18	5,45
Kansas	67,6	66,8	67	67,1	2264	5,66	4,17	5,53
Kentucky	62,7	65,5	66,4	65,5	5949	5,65	4,18	5,51
Louisiana	64,7	65,5	66,2	65,2	9065	5,54	4,17	5,45
Maine	67,3	66,5	66,7	67,1	531	5,73	4,19	5,58
Maryland	68	66,1	66,2	66,1	16773	5,56	4,16	5,46
Massachusetts	68,1	66,8	66,1	66,8	8164	5,64	4,17	5,51
Michigan	65,6	65,9	65,8	65,7	15080	5,61	4,17	5,49
Minnesota	68,9	66,7	66,6	67,2	4468	5,68	4,18	5,54
Mississippi	63,6	64,9	66,3	64,5	7927	5,56	4,15	5,46
Missouri	65,5	66	66,2	65,8	6222	5,59	4,17	5,48
Montana	68,5	67,7	67	67,7	177	5,53	4,25	5,48
Nebraska	68,5	66,8	67	66,7	1320	5,73	4,19	5,57
Nevada	65,2	67	66,4	67	3972	5,68	4,19	5,54
New Hampshire	68,4	67,7	67,7	68,3	671	5,68	4,16	5,55
New Jersey	66,1	66,2	66,1	66,2	16286	5,62	4,18	5,49
New Mexico	66,7	67	66,8	67	1388	5,67	4,17	5,55
New York	66,2	67,1	66,1	67,3	20827	5,64	4,16	5,51
North Carolina	65,7	65,5	65,8	65,4	17941	5,61	4,16	5,49
North Dakota	67,4	67,3	66,7	67,2	539	5,57	4,18	5,49
Ohio	64,6	65,9	66,2	66	19518	5,61	4,18	5,5
Oklahoma	65,2	66	65,8	65,7	3115	5,61	4,18	5,5
Oregon	67,1	67,1	66,8	67,3	2123	5,7	4,18	5,56
Pennsylvania	66,5	66,1	66,1	66,2	19510	5,62	4,18	5,5
Rhode Island	65,5	65,9	66,8	65,7	1116	5,66	4,17	5,5
South Carolina	65,2	65,5	66,2	65,3	9290	5,61	4,16	5,49
South Dakota	68	68	66,6	67,8	594	5,56	4,16	5,46
Tennessee	64	65,9	66,3	65,6	7918	5,62	4,18	5,49
Texas	66,6	65,9	65,8	66,1	43271	5,6	4,17	5,48
Utah	68,8	67,4	67,2	67,3	1842	5,65	4,18	5,53
Vermont	68,6	67,1	68	67,6	134	5,87	4,19	5,66
Virginia	67,7	66,5	66,2	66,2	17672	5,61	4,16	5,49
Washington	67,7	66,9	66,9	66,7	3525	5,71	4,17	5,55
West Virginia	61,3	66	66,4	65,6	1767	5,67	4,19	5,53
Wisconsin	67,3	67,3	66,1	67	5111	5,68	4,15	5,51
Wyoming	67,9	67,4	67,1	68,2	241	5,63	4,2	5,51
Max	69,7	68	68	68,3	43271	5,87	4,25	5,66
Min	61,3	64,9	65,7	64,5	134	5,49	4,15	5,45
Average	66,5	66,5	66,5	66,5	8867,08	5,64	4,18	5,51

Table 5.11: Values for every continental US State (excluding Alaska), when considering the Gallup-Healthways score, the well-being score predictions from Models *A, B* and *C*, the number of Tweets available, and the predicted average Valence, Arousal and Dominance scores.

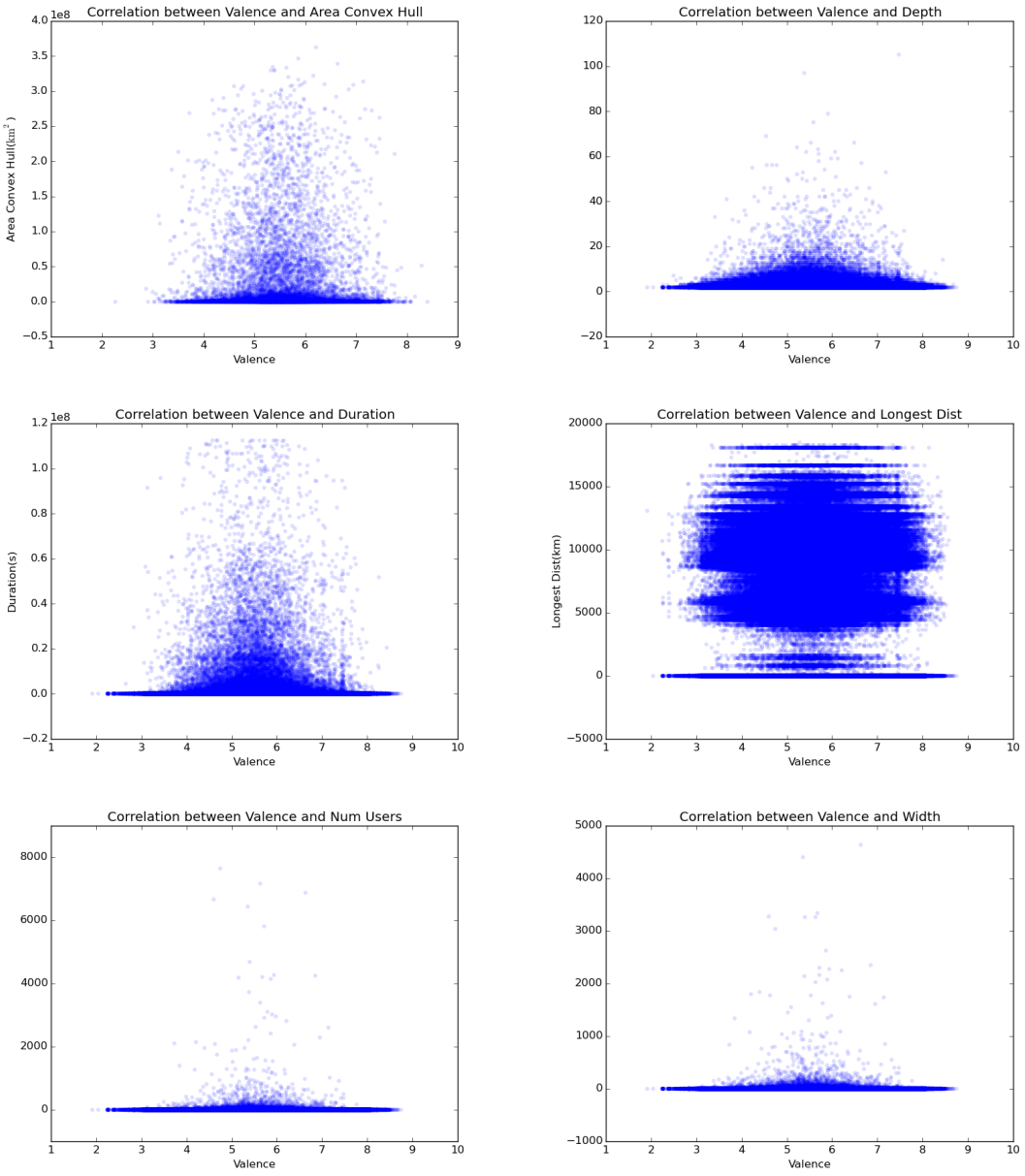

Figure 5.9: Correlation between Valence and the Diffusion metrics.

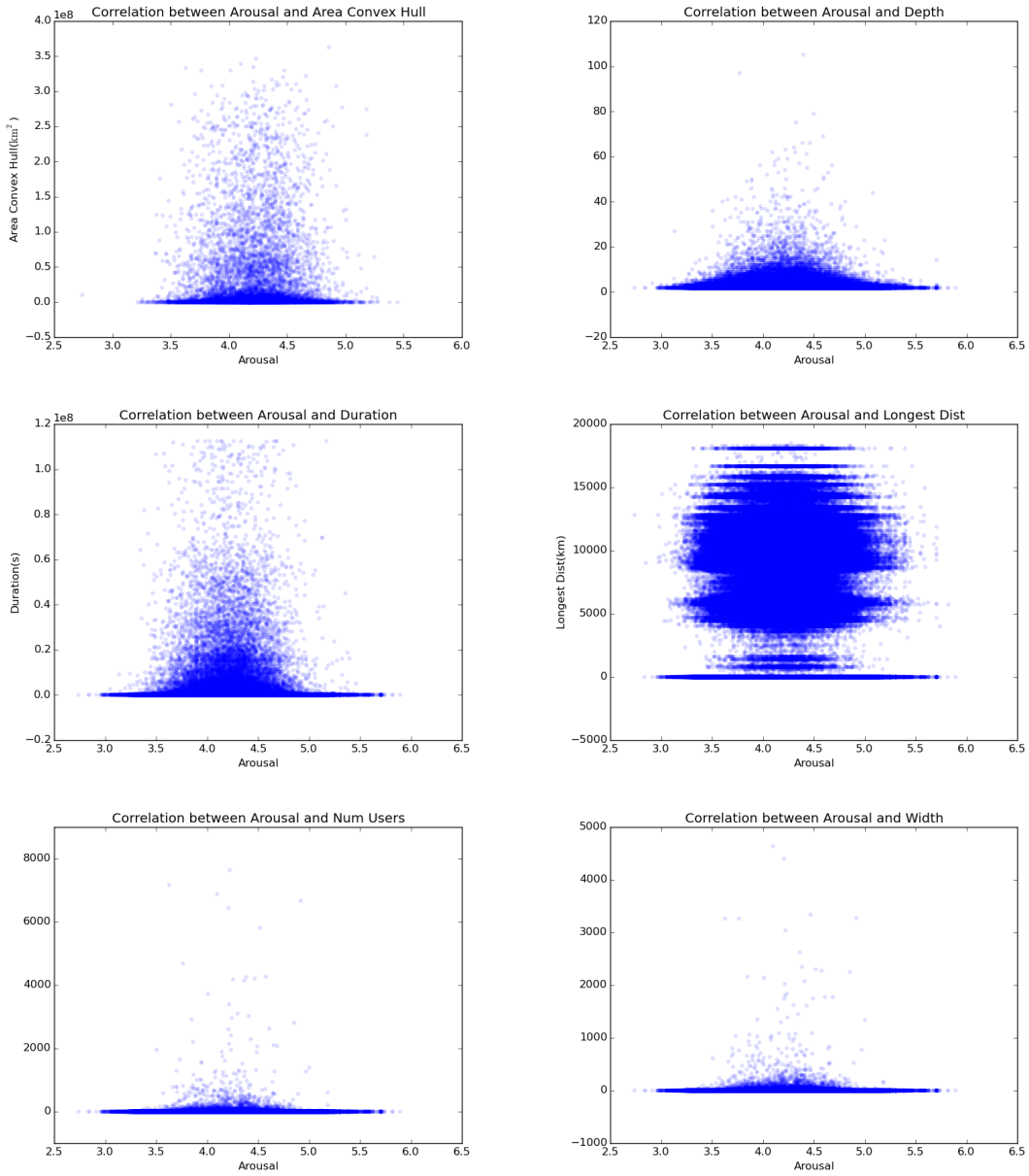

Figure 5.10: Correlation between Arousal and the Diffusion metrics.

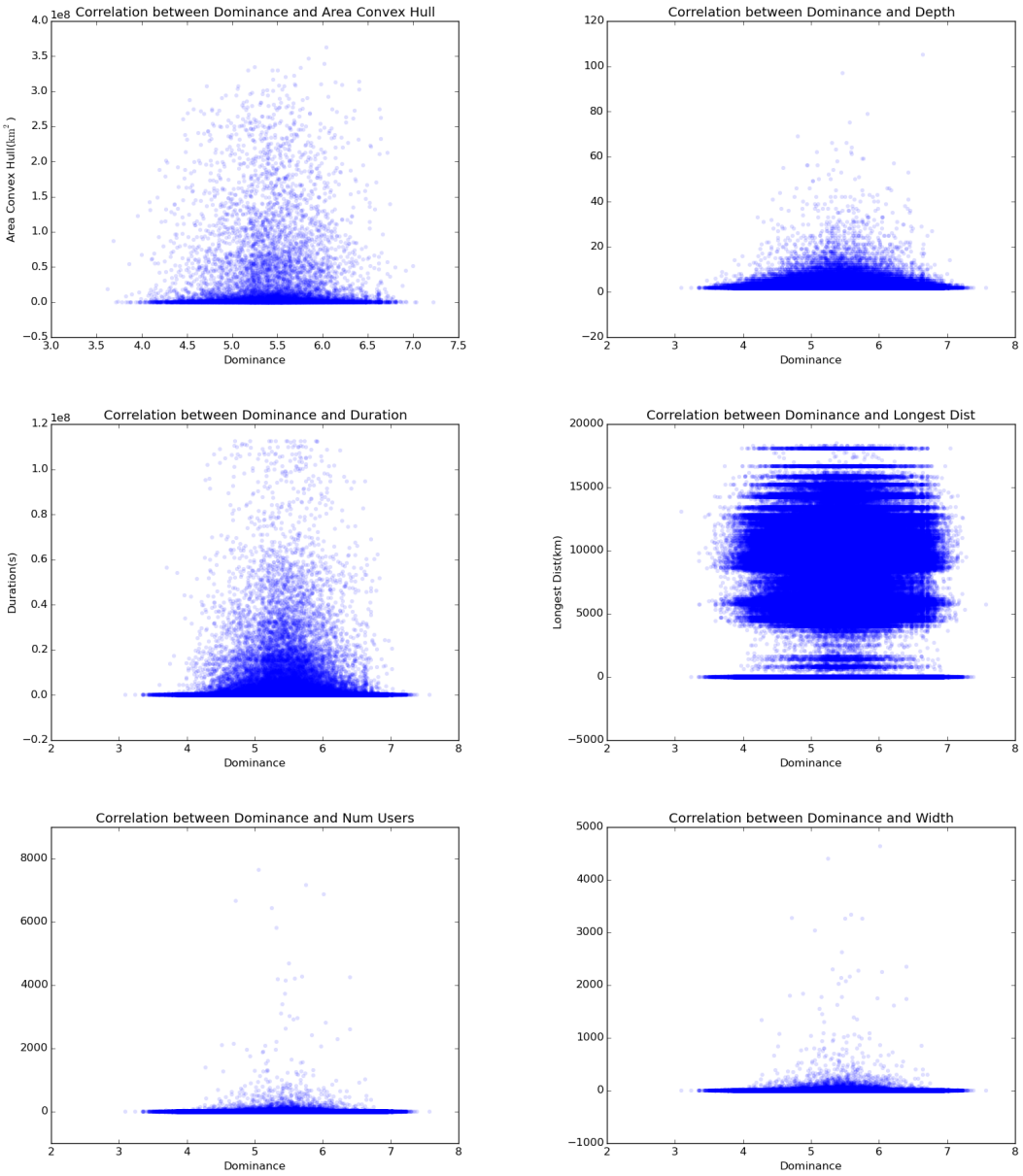

Figure 5.11: Correlation between Dominance and the Diffusion metrics.

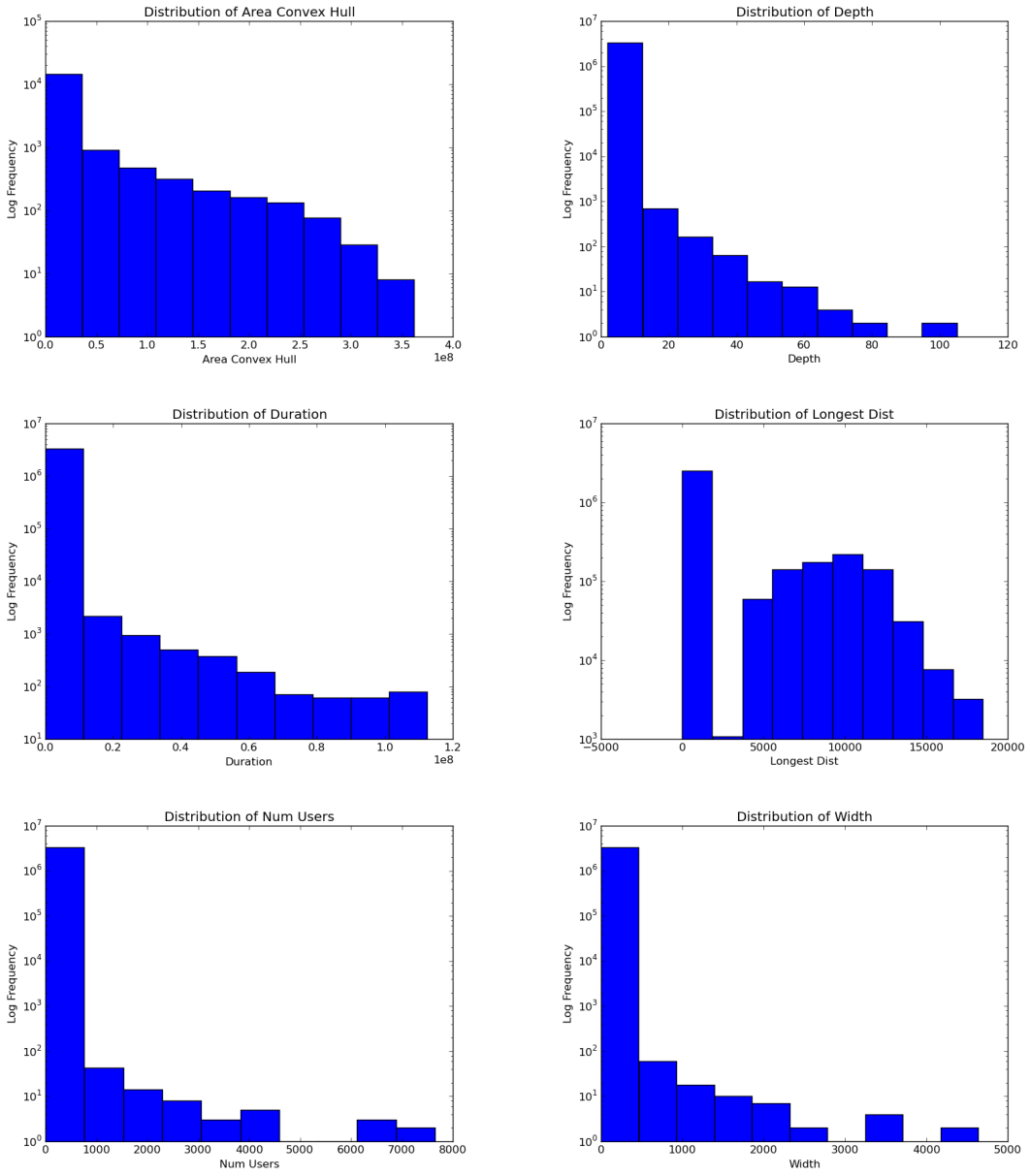

Figure 5.12: Distribution for each one of the considered Diffusion Metrics on a logarithmic scale.

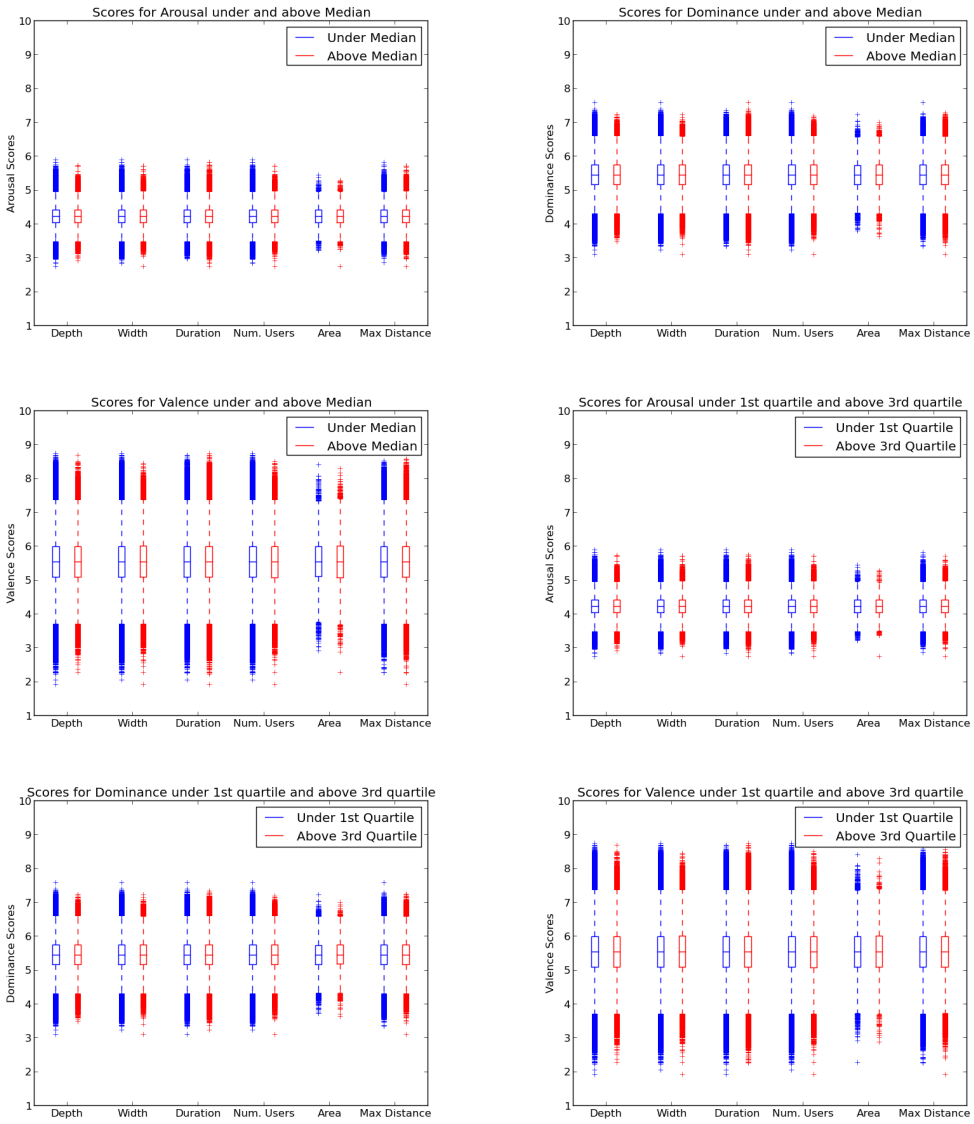

Figure 5.13: Distribution of scores for all the considered diffusion metrics above the median/first quartile and under the median/third quartile for Valence, Arousal and Dominance.

Chapter 6

Conclusions

This chapter summarizes the work described in the present dissertation, focusing on enumerating the set of contributions, as well as pointing some directions for future research based on the current achievements and results.

6.1 Summary of Contributions

The most valuable contribution and the focus of the research in this work was the building and evaluation of a method for predicting emotional scores, in terms of valence, arousal and dominance for words and short texts based on neural embeddings. The method consisted in building regression models (i.e., leveraging on Random Forests, k-NN, and Kernel Ridge) that based on neural embeddings of words or short texts would predict emotional scores in the aforementioned emotional dimensions. These models were trained leveraging on emotional lexicons (i.e., dictionaries that associate words to emotional scores) such as ANEW (Bradley & Lang, 1999) and the lexicon from Warriner *et al.* (2013a), and on the neural embeddings associated to the words present in the aforementioned emotional lexicons. Then, having these regression models and embeddings associated to a given word/short text, the method is able to predict emotional scores, in terms of valence, arousal and dominance to this same word/short text. This method produced state-of-the-art results when predicting scores for words in English, i.e., the agreement between the predictions and the scores from emotional lexicons are in line and sometimes higher than the agreement between humans when evaluating words in terms of emotional scores. This same method was also applied in a cross lingual approach, i.e., the models were trained with

emotional lexicons in English and used to predict emotional in other languages (Spanish, Portuguese, Italian and German). For this to be possible the embeddings used to train and to predict, i.e., embeddings for English words and embeddings for words in a foreign language had to be in the same vector space. To do so, and taking inspiration from Faruqui & Dyer (2014), CCA was used to transpose the two sets of embeddings to the same vector space. When evaluating this last approach (predicting scores for words in foreign languages), the results were not as high when predicting scores for words in English, but significant correlations were measured. Is also worth to mention, that these last two approaches, were later used to produce extended emotional lexicons both for English words and for words in Spanish, Portuguese, Italian and German for all the words present in the embeddings models leveraged on this work. The previous existent lexicons for these languages (Bradley & Lang, 1999; Montefinese *et al.*, 2014; Redondo *et al.*, 2007; Schmidtke *et al.*, 2014; Soares *et al.*, 2012; Warriner *et al.*, 2013a) since being annotated by humans were very limited in terms of the number of words they contained (13,915 words in the case of the lexicon from Warriner *et al.* (2013a)). Finally, and following the same line of thoughts, regressions models capable of predicting emotional scores for short texts were also built. Having embeddings both for words in English and for sentences present in datasets of short texts annotated according to emotional dimensions, regression model were trained using words from Warriner *et al.* (2013a), and used to predict the scores associated to these same texts, based on their embeddings. The results obtained when evaluating this last approach, were under the results obtained when predicting scores for words, besides having measured significant correlation when predicting scores for some datasets. This suggests there is still space for improvement.

Finally, this work also contributed with two different applications which leverage on the method for predicting emotional scores associated to short texts. Both the applications resort to a dataset composed of tweets. The first application, based on the embeddings and scores predicted for each one of tweets issued from the continental USA states (excluding Alaska), predicts the well-being score for each one of these states. The quality of the predictions were found to be more robust than previous comparable studies. The second application, consisting in trying to see if there is a correlation between the score predicted to a given tweet, and the way this same tweet spreads in Twitter according some diffusion metrics (e.g., the number of retweets, the number of user the tweet reaches, etc). The results were found to be inconclusive, and no evident relationships between emotional dimensions and diffusion metrics were found.

6.2 Future Work

Despite the interesting results, there are also several possible paths for improvement on the various facets of this work. Regarding the method for predicting emotional scores for words and short texts there are a set of experiments that would be interesting to perform:

- Experimenting with alternative word embeddings procedures (Liu *et al.*, 2015; Pennington *et al.*, 2014), since the aforementioned described methods are not design dependent on the type of embeddings which are used;
- Testing the method for inferring emotional scores for short texts with datasets of bigger size;
- Extending the methodology for predicting emotional scores for short texts to other languages than English, by following a similar approach used on the method for predicting scores for words in other languages, i.e., resorting to Canonical Correlations Analysis (CCA) for leveraging English data with the purpose of estimating embeddings for other languages, to then being able to use regression models to predict scores for short texts in these languages;
- Besides the good results when experimenting with the usage of CCA with the purpose of transforming embeddings from different languages to the same vector space, this particular technique is only able to reveal linear relationships. Would be interesting to experiment with methods introduced by Andrew *et al.* (2013); Lopez-Paz *et al.* (2014) for extending CCA in order to learn complex nonlinear relations between data.

Regarding the two application presented in Chapter 5, there are some interesting experiments worth to try:

- Applying the same methodology to predict other measurable characteristics of regions and populations (e.g., GDP, population size, etc);
- Testing the methodology used to reconstruct the diffusion path for a given tweet with datasets of similar social networks where this diffusion path is explicit. Until the time this document is being written, it was impossible to have acess datasets of that kind;
- Verifying if other semantic characteristics beside the emotional scores studied on this word influence the diffusion process of a given tweet.

Bibliography

ANDREW, G., ARORA, R., BILMES, J. & LIVESCU, K. (2013). Deep canonical correlation analysis. In *Proceedings of the International Conference on Machine Learning*.

BARONI, M., DINU, G. & KRUSZEWSKI, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

BERGSMA, S., LIN, D. & GOEBEL, R. (2009). Web-scale n-gram models for lexical disambiguation. In *Proceedings of the International Joint Conferences on Artificial Intelligence*.

BESTGEN, Y. & VINCZE, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods*, **44**.

BLEI, D.M., NG, A.Y. & JORDAN, M.I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, **3**.

BLISS, C.A., KLOUMANN, I.M., HARRIS, K.D., DANFORTH, C.M. & DODDS, P.S. (2012). Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal of Computational Science*, **3**.

BRADLEY, M.M. & LANG, P.J. (1999). Affective norms for english words (ANEW): Stimuli, instruction manual and affective ratings. Tech. Rep. C-1.

BRADLEY, M.M. & LANG, P.J. (2007). Affective norms for english text (ANET): Affective ratings of text and instruction manual. Tech. Rep. D-1.

BRADLEY, M.M. & LANG, P.J. (2010). Affective norms for english words (ANEW): Affective ratings of words and instruction manual. Tech. Rep. C-2.

BRADLEY, M.M., LANG, P.J., BRADLEY, M.M. & LANG, P.J. (1999). Affective norms for english words : Instruction manual and affective ratings.

BREIMAN, L. (2001). Random forests. *Machine Learning*, **45**.

BREIMAN, L., FRIEDMAN, J., STONE, C. & OLSHEN, R. (1984). *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series, Taylor & Francis.

CALVO, R.A. & D'MELLO, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, **1**.

CHA, M., HADDADI, H., BENEVENUTO, F. & GUMMADI, P.K. (2010). Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of the International AAAI Conference on Web and Social Media*.

CHRISTAKIS, N.A. & FOWLER, J.H. (2013). Social contagion theory: examining dynamic social networks and human behavior. *Statistics in Medicine*, **32**.

DE MELO, G. & WEIKUM, G. (2009). Towards a universal WordNet by learning from combined evidence. In *Proceedings of the ACM Conference on Information and Knowledge Management*.

DE MELO, G. & WEIKUM, G. (2010). Untangling the cross-lingual link structure of Wikipedia. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

DODDS, P.S., HARRIS, K.D., ISABEL M. KLOUMANN, C.A.B. & DANFORTH, C.M. (2011a). Temporal patterns of hapiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE*, **6**.

DODDS, P.S., HARRIS, K.D., KLOUMANN, I.M., BLISS, C.A. & DANFORTH, C.M. (2011b). Temporal patterns of happiness and information in a global social network: hedonometrics and twitter. *PLoS ONE*, **6**.

EICHSTAEDT, J.C., SCHWARTZ, H.A., KERN, M.L., PARK, G., LABARTHE, D.R., MERCHANT, R.M., JHA, S., AGRAWAL, M., DZIURZYNSKI, L.A., SAP, M., WEEG, C., LARSON, E.E., UNGAR, L.H. & SELIGMAN, M.E. (2015). Psychological language on twitter predicts county-level heart disease mortality. *Psychological Science*, **26**.

FARUQUI, M. & DYER, C. (2014). Improving vector space word representations using multilingual correlation. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*.

FRANCISCO, V., HERVÁS, R., PEINADO, F. & GERVÁS, P. (2012). Emotales: creating a corpus of folk tales with emotional annotations. *Language Resources and Evaluation*, **46**.

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, **33**.

GHOSH, R. & LERMAN, K. (2011). A framework for quantitative analysis of cascades on networks. In *Proceedings of the ACM International Conference on Web Search and Data Mining*.

GILL, A.J., NOWSON, S. & OBERLANDER, J. (2009). What are they blogging about? personality, topic and motivation in blogs. In *Proceedings of the International AAAI Conference on Web and Social Media*.

GOLDBERG, Y. & LEVY, O. (2013). word2vec explained: deriving mikolov et al.'s negative- sampling word-embedding method. *arXiv*.

GUERINI, M. & STAIANO, J. (2015). Deep feelings: A massive cross-lingual study on the relation between emotions and virality. In *Proceedings of the International Conference on World Wide Web*.

GUILLE, A., HACID, H., FAVRE, C. & ZIGHED, D.A. (2013). Information diffusion in online social networks: A survey. *SIGMOD Record*, **42**.

HASTIE, T.J. & TIBSHIRANI, R.J. (1990). *Generalized additive models*. CRC Press.

KAMATH, K.Y., CAVERLEE, J., CHENG, Z. & SUI, D.Z. (2012). Spatial influence vs. community influence: Modeling the global spread of social media. In *Proceedings of the ACM Conference on Information and Knowledge Management*.

KAMATH, K.Y., CAVERLEE, J., LEE, K. & CHENG, Z. (2013). Spatio-temporal dynamics of online memes: A study of geo-tagged tweets. In *Proceedings of the World Wide Web Conference*.

KRAMER, A.D. (2010). An unobtrusive behavioral model of gross national happiness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

LAFFERTY, J., MCCALLUM, A. & PEREIRA, F.C. (2001). Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning*.

LAWLESS, N.M. & LUCAS, R.E. (2011). Predictors of regional well-being: A county level analysis. *Social Indicators Research*, **101**.

LE, Q. & MIKOLOV, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the International Conference on Machine Learning*.

LERMAN, K., GHOSH, R. & SURACHAWALA, T. (2012). Social contagion: An empirical study of information spread on Digg and Twitter follower graphs. *CoRR*.

LESKOVEC, J. (2011). Social media analytics: Tracking, modeling and predicting the flow of information through networks. In *Proceedings of the International Conference on World Wide Web*.

LI, J., WANG, X. & HOVY, E. (2014). What a nasty day: Exploring mood-weather relationship from Twitter. In *Proceedings of the ACM Conference on Information and Knowledge Management*.

LIU, Y., LIU, Z., CHUA, T.S. & SUN, M. (2015). Topical word embeddings. In *Proccedings of the AAAI Conference on Artificial Intelligence*.

LOFF, J., REIS, M. & MARTINS, B. (2015). Predicting well-being with geo-referenced data collected from social media platforms. In *Proceedings of the ACM/SIGAPP Symposium On Applied Computing*.

LOPEZ-PAZ, D., SRA, S., SMOLA, A., GHAHRAMANI, Z. & SCHÖLKOPF, B. (2014). Randomized nonlinear component analysis. In *Proceedings of the International Conference on Machine Learning*.

MANDERAA, P., KEULEERSA, E. & BRYSBAERTA, M. (2015). How useful are corpus-based methods for extrapolating psycholinguistic variables? *The Quarterly Journal of Experimental Psychology*.

MIKOLOV, T., CHEN, K., CORRADO, G. & DEAN, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the Workshop at the International Conference on Learning Representations*.

MONTEFINESE, M., AMBROSINI, E., FAIRFIELD, B. & MAMMARELLA, N. (2014). The adaptation of the affective norms for english words (ANEW) for italian. *Behavior Research Methods*, **46**.

MURPHY, S. (2014). Validation of the affective norms for english words (ANEW) as weighted computerized language measures of arousal, valence and dominance. Tech. rep., Pacella Research Center of the New York Psychoanalytic Society and Institute.

OWOPUTI, O., O'CONNOR, B., DYER, C., GIMPEL, K., SCHNEIDER, N. & SMITH, N.A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing*.

PALTOGLOU, G. & THELWALL, M. (2013). Seeing stars of valence and arousal in blog posts. *IEEE Transactions on Affective Computing*, **4**.

PALTOGLOU, G., THEUNIS, M., KAPPAS, A. & THELWALL, M. (2013). Predicting emotional responses to long informal text. *IEEE Transactions on Affective Computing*, **4**.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M. & DUCHESNAY, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, **12**.

PENNEBAKER, J.W., FRANCIS, M.E. & BOOTH, R.J. (2001). Linguistic inquiry and word count. *Mahway: Lawrence Erlbaum Associates*, **71**.

PENNINGTON, J., SOCHER, R. & MANNING, C.D. (2014). Glove: Global vectors for word representation. *Proceedings of the Conference on Empiricial Methods in Natural Language Processing*.

POLLOCK, V., CHO, D.W., REKER, D. & VOLAVKA, J. (1979). Profile of mood states: the factors and their physiological correlates. *The Journal of Nervous and Mental Disease*, **167**.

QUERCIA, D., ELLIS, J., CAPRA, L. & CROWCROFT, J. (2011). In the mood for being influential on Twitter. In *Proceedings of the IEEE International Conference on Social Computing*.

RECCHIAA, G. & LOUWERSE, M.M. (2014). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *The Quarterly Journal of Experimental Psychology*, **68**.

REDONDO, J., FRAGA, I., PADRÓN, I. & COMESANA, M. (2007). The spanish adaptation of ANEW (affective norms for english words). *Behavior Research Methods*, **39**.

RITTER, A., ETZIONI, O., CLARK, S. & MAUSAM (2012). Open domain event extraction from Twitter. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*.

ROMERO, D.M., GALUBA, W., ASUR, S. & HUBERMAN, B.A. (2010). Influence and passivity in social media. In *Proceedings of the International Conference on World Wide Web*.

ROMERO, D.M., MEEDER, B. & KLEINBERG, J. (2011). Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of the International Conference on World Wide Web*.

RUSSEL, J.A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, **6**.

SADIKOV, E., MEDINA, M., LESKOVEC, J. & GARCIA-MOLINA, H. (2011). Correcting for missing data in information cascades. In *Proceedings of the ACM International Conference on Web Search and Data Mining*.

SCHMIDTKE, D.S., SCHRÖDER, T., JACOBS, A.M. & CONRAD, M. (2014). ANGST: Affective norms for german sentiment terms, derived from the affective norms for english words. *Behavior Research Methods*, **46**.

SCHWARTZ, H.A., EICHSTAEDT, J.C., DZIURZYNSKI, L., KERN, M.L., SELIGMAN, M.E., UNGAR, L.H., BLANCO, E., KOSINSKI, M. & STILLWELL, D. (2013a). Toward personality insights from language exploration in social media. In *Proceedings of the AAAI Workshop on Analyzing Microtext*.

SCHWARTZ, H.A., EICHSTAEDT, J.C., KERN, M.L., DZIURZYNSKI, L., LUCAS, R.E., AGRAWAL, M., PARK, G.J., LAKSHMIKANTH, S.K., JHA, S., SELIGMAN, M.E. & UNGAR, L. (2013b). Characterizing geographic variation in well-being using tweets. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

SELIGMAN, M.E. (2011). *Flourish: A visionary new understanding of happiness and well-being.*. Simon and Schuster.

SOARES, A.P., COMESAÑA, M., PINHEIRO, A.P., SIMÕES, A. & FRADE, C.S. (2012). The adaptation of the affective norms for english words (ANEW) for european portuguese. *Behavior Research Methods*, **44**.

TAXIDOU, I. & FISCHER, P.M. (2014). Online analysis of information diffusion in Twitter. In *Proceedings of the International Conference on World Wide Web*.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **58**.

WARRINER, A., KUPERMAN, V. & BRYSBAERT, M. (2013a). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, **45**.

WARRINER, A., KUPERMAN, V. & BRYSBAERT, M. (2013b). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, **45**.

WOOD, S.N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**.

ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**.