# Diary mining: predicting emotion from activities, people and places

A thesis submitted in partial fulfillment of the requirement for the degree of Doctor of Philosophy

**Shahd Alahdal**

**January 2020**

**Cardiff University**
**Department of Computer Science and Informatics**

# *Acknowledgements*

I want to thank my academic supervisor, Professor Stuart Allen, for his continuous support and guidance, for the valuable times and discussions, for turning my Ph.D. experience into an extraordinary one. I also want to thank Dr. Coral Walker and Dr. Alia Abdelmoty. I would like to extend my thanks to my sponsors: King Abdulaziz University, the University of Jeddah and Saudi Cultural Berau, for their financial support.

I am grateful and thankful to my daughters, Noor, my life, and Malak, my happiness. Thanks won't be enough for my husband, Mishal, for his patience and support. My deep and sincere gratitude to, my Mother, Sisters, and Brother for always being there for me and for their priceless support. And Dad, you would have made a lot of fun at me now, wish you haven't left in the middle of this.

# Abstract

Diary methods are concerned with collecting qualitative information from people about their everyday lives and are commonly used in many fields such as psychology, sociology and medicine to understand human behaviour and improve mental health. By its nature, the data is difficult to analyse and time-consuming to process manually, creating a gap between collection, analysis and intervention. Technologies such as machine learning have the potential to shrink this gap, save time and effort, and hence give deeper insight into the diary data.

Computer science technologies have been heavily used by many disciplines to understand humans. One such application is emotion detection from text, which is the process of automatically identifying the emotion that is either directly expressed by the author or the underlying emotion that prompted the author to write a text. Studies have shown promising results using different features extracted, whether linguistic or others (e.g., number of followers). However, very few have used activities for emotion prediction from text, and none of these have combined activities with other associated situational features from the relevant event.

The research in this thesis proposes an approach to predict emotion from self-recorded personal textual diaries using a small set of domain-specific features. Daily activities, in association with people and places, are used as the main indicators of an individual's current situation. The association of these factors with emotion has been well-studied independently in psychology, which has motivated this investigation to validate the combination of all three features and test their ability to predict emotion from a computer science perspective.

This research begins by proposing a framework to classify short diary entries into a small number of high-level personal activities (work/study, social/family, food/drink, leisure, essentials) and represents them as low dimensional probability vectors using unsupervised (clustering) and supervised (classification) machine learning techniques. In view of the fact that these entries are characterised by sparseness, and that there is lack of training data as they are highly personal, this framework applies a transfer learning approach by exploiting previously acquired knowledge as a foundation step, using a pre-trained word embedding model on similar, but not identical, and easily obtained publicly available data (tweets). Furthermore, references to people and places are also recognised from the text using information extraction techniques.

These automatically extracted features are then used for predicting emotion, utilising different emotion schemes, including Ekman's basic emotion model, the Circumplex model, together with simpler classification into pleasantness/unpleasantness, and emotional/neutral states. In addition, different learning strategies for predicting emotion are compared, including the use of personalised and global training data. This research has shown that activities, people, and places can successfully predict some emotions from the text, especially 'happiness' and 'neutral'.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Emotion plays an important role in many aspects of people's lives, having a powerful effect on their health, work, learning, economic behaviour, decision making and communication. For example, there is a strong relationship between the increase in positive emotions and creativity of employees at work [1]. Research has shown that being in a happy state of mind enhances learning, facilitates students' understanding and efficient retrieval of information [2]. It has also been shown that positive emotions experienced by online shoppers may lead to impulse purchases [3] and that experiencing a wide range of positive emotions reduces the risk of diseases [4]. On the other hand, it has been suggested that depression, a negative emotion, can contribute to cancer progression and make treatment less effective [5]. An individual's emotions may be influenced by various factors. For example, happiness is enhanced in several ways: by engaging in regular exercise; having enough sleep; maintaining strong friendships and socialising more frequently with supportive people [6]. Indeed, many other factors have been shown to affect emotions positively or negatively, such as the weather [7], music [8], diet [9], sleep [10], personal relationships and places [11].

For many years, the study of emotions has been the focus of attention of researchers in psychological science. With the beginning of the twenty-first century, it has begun to attract the interest of researchers in other sciences including neuroscience, psychiatry, biology, genetics, behavioural economics, computer science and many others, eventually growing to become a field of its own: *affective science* [12]. Each scientific field contributes differently to the study of emotion. Some look at it as a neural mechanism, others study the physiological responses, while others investigate it as a cognitive activity. Studies vary tremendously, ranging from how chocolate is associated with pleasure [13] to how depression is the most important risk factor for lifetime suicide [14].

More recently, people started to use social media platforms (e.g., the Facebook mood manipulator) and many other applications (e.g., mood tracker, easy mood diary) to track and share their emotions. This has provided the opportunity to collect data from large populations and has made it easier to understand emotions in non-laboratory settings. Consequently, the study of emotions has become a growing sub-field in computer science. Known as *affective computing*, it aims to enable intelligent systems to recognise, infer, express, interpret and respond to human emotions [15]. It is concerned with improving human-computer interaction, for example, by reducing and handling user frustration once it has occurred [16], as well as dealing with human-human interaction, for example, deception detection by processing facial emotions, especially those mismatches between facial and verbal messages [17]. It also serves many fields such as business and marketing, in tasks such as predicting customers' attitudes towards products and building recommendation systems [18]. Many researchers working in the area of e-mental health build predictive models to understand and explore the relationships between concepts and emotions using digital data (e.g., diary data, facial expressions, speech samples, smartphone measures, GPS movement data and many others), which helps with health monitoring, treatment selection and treatment personalisation [19, 20]. Much work has been conducted to detect a human's emotional state focusing on different communication channels, such as audio [21], facial [22], gesture [23], eye gaze [24] and text [25].

There is a general agreement that there are links between features of the physical environment/associated situations and emotion. Interactionist theories argue that people attempt or desire to be in an environment which satisfies their psychological needs [26], implying that the changes between positive and negative emotions that people typically encounter in their everyday lives mainly occur as a result of situational cues [11] which can be represented by people's interactions or their company, objects, events, activities and places [27].

In today's digital world, with the advent of technology, people are extensively using different modes to express themselves, such as written, visual and quantitative [28]. For example, by writing about their lives in blogs, tweets and status updates on Facebook, taking selfies and posting them to Instagram and some wear activity trackers on their wrists to record their habits (e.g., steps, heart rate); all of which provides a greater opportunity to collect data on people's lives, feelings and attitudes.

Diaries are one important source of situational information that people use to preserve the events and thoughts in their daily lives. In diaries, people report how they spend their time, what they do, they also note the important people in their lives, places, emotions and much more. Therefore, for several years, diaries have been the focus of attention of researchers in different fields such as psychology, social science and medicine. They use "diary methods" to understand and analyse different human phenomena such as relationships, social interactions and emotions [29, 30]. Nowadays, diaries are taking more digital, shorter forms, and a variety of applications have emerged for easily recording and retrieval, as well as blogs, micro-blogs and social media websites where people post their day-to-day life events.

Electronic diaries (e-diaries) are a feasible method for assisting in the analysis of emotional experiences and capturing a range of detailed and interesting qualitative and quantitative data about people's moods, stresses, responses and general functions [31, 32]. For example, e-diaries which require emotional experience reporting might encourage emotional processing and facilitate a reduction in symptoms of anxiety [33]. Moreover, components of emotion regulation as captured by e-diaries predict important health outcomes like daily pain and function in children with certain medical conditions [34].

Recorded moments of people's lives, and the increasingly available human-generated data originating on the web (e.g., social media, news) have helped to overcome the limitations of available data and gives the opportunity to researchers in different fields to utilise this continuous, large stream of data [35]. In line with this, text mining has become an important, and trendy, field. It is concerned with discovering meaningful information from unstructured or semi-structured text that were previously unknown. It widely covers a large set of related algorithms and approaches including: information extraction; natural language processing; text summarisation; topic modelling; clustering and classification. Selection of the appropriate technique for mining text reduces the time and effort to find relevant patterns [36–39]. Text mining has many applications in different fields including medicine, genetics, social media, education, psychology and many others. "In principle, any technology that can help people change their mindset and behaviour can be used to improve psychological well-being" [40].

## 1.1 Research problem and motivation

Research in psychology has intensively investigated the many factors that can affect emotion, including activities, people and places independently in different studies. In recent years, a growing number of studies from computer science have proposed techniques to predict emotion. However, very few studies have used activities to predict emotions, and none of these considered combining activities with other factors from the relevant event (i.e. people and places) to predict emotion.

The research in this thesis provides a framework to investigate the effect or association between different external environmental features, namely activities, people and places, on individuals' emotion. In particular, we consider the problem of combining these three factors as digital features to improve emotion prediction.

Based on this insight, and from the interest of investigating and validating psychological theories via computers, this thesis has two main building blocks. Firstly, we consider activity classification in isolation, intending to infer daily activities from short text using semi-supervised machine learning algorithms, exploit the external data resource as a foundation or a starting point. Secondly, these activity vectors are used in association with other features: people and places, that are automatically extracted from the text for a classification task to predict emotion. Such methods are quite prevalent among practitioners and researchers. For example, researchers in mental health use predictive models to explore the relationships between concepts and emotions aiming to understand the influence of various factors on emotions, detect abnormal emotional state, and help improve mental health.

## 1.2 Aims and objectives

The research in this thesis aims to assess the capability of machine learning approaches to detect activities, people, and places from short texts and combine them as features to predict emotion. The objectives of this research are as follows:

- To provide effective automated methods to analyse diary data that would be suitable to support, for example, psychologists/medical practitioners monitoring their patients' environment to identify causes of relaxation/anxiety.
- To develop a simple and effective framework to categorise daily activities that can enhance our understanding of humans in non-laboratory settings.
- To develop a prediction model for emotion using a small set of specific domain features (activities, people and places) instead of direct linguistic references (i.e., emotion words).
- To apply different learning strategies in emotion prediction and evaluate prediction modes from the perspective of different emotion classification schemes commonly used in psychology.

## 1.3 Research questions

- **Research Question 1**: *Emotion extraction: how can we use the activity, people, and places specified in short diary entries to extract or predict emotion utilising personal and global data?*

Unlike existing approaches for predicting emotion from short texts, here a small set of domain-specific features are tested for their ability to indicate some emotions that people are experiencing in their current situation. To reduce the scope of features, we first consider activities in isolation:

- **Research Question 2**: *Daily activity categorisation: how can we classify short diary entries into a small number of high level activity categories?*

The challenge lies in having short unstructured diary text, which is characterised by being sparse, where there is a lack of a sufficient co-occurrence of words and context information that help to infer semantics. Another challenge is the lack of training data for diaries as they are highly personal and hard to collect. These two challenges raised a question: Will using an external domain as a source for training data boost the classification task?

## 1.4 Research hypothesis and methodology

**This research addresses the following main hypothesis:**
*Activities, people and places extracted from diaries are three situational factors that can together improve emotion classification from short texts.*
**The methodology of the research follows the following steps:**

- **Collection** of a representative data set of personal human-generated diary over a period of time, as text entries that describe their current situation and associated emotion labels.
- **Development** of a framework that uses publicly available data from an external domain to detect and categorise activities, and combines these with people and places to predict emotion.
- **Analysis and evaluation** across different representations of emotion, using crowdsourced ground truth and standard measures of evaluation (e.g., precision, recall and F-measure).

## 1.5 Research contributions

The contribution of this research lies in building a predictive model for emotions in context: activities, people, and places, which were not previously combined, taken from short diary textual entries by applying supervised and unsupervised machine learning techniques. In answering the research questions, the main contributions made in this work are outlined below:

- Proposal and evaluation of an emotion prediction approach using a small set of domain-specific features, mainly daily activities, in association with people and places as the context's presenters of individuals in real life. Using well-studied techniques in the machine learning field and drawing on the research into human emotions from psychology, the research automatically extracts features from diaries and infers relevant emotions.

- Proposal and evaluation of a framework based on transfer learning and clustering, which can accurately classify the relevant activity to a short diary entry. This approach has overcome the challenges of a lack of large scale training data by using an external data resource and neural network.

- Proposal and evaluation of different approaches to handle trivial diary entries in the activity classification task. This work introduces different processing and analysis methods and assesses their effect of the nature of the activities on the classification results.

- Investigation of the effect of different emotional models as well as different learning strategies for predicting emotions, showing that the origin and structure of the models influence the emotion classification task.

## 1.6 Thesis structure

The remaining chapters of the thesis are organized as follows:
**Chapter 2** - *Background and related work* - provides an overview of a diary and the diary methods. Diary features such activities, people, places and emotions. As well as text mining techniques and introduces the data collection methods used in this work.
**Chapter 3** - *Activity categorisation framework* - describes the proposed model for classifying diary entries in terms of daily activities.
**Chapter 4** - *Classification analysis* - provides an evaluation of the daily activity classification framework; it reports the results, and provides different views and levels of analysis.
**Chapter 5** - *Emotion prediction* - looks at different aspects of emotions prediction. Comparing different machine learning strategies for training and testing the classifiers: personal and global. Comparing the emotion prediction performance of different emotion schemes.
**Chapter 6** - *Conclusions, limitations and future work* - concludes the thesis by summing the major contributions of this research and suggesting possible directions for future research.

# Chapter 2

# Background and related work

## 2.1 Introduction

Writing diaries is an important activity for many people in order to preserve precious memories. Diaries these days are increasingly generated digitally, due to the ease of saving, organising, retrieval and, most importantly, the ubiquitous availability of the required technology; all of which has encouraged them to shift from physical forms. Consequently, an enormous amount of applications have become available for users to record their diaries, as well as a number of different platforms that can be used for posting and sharing their daily events. These include blogs, microblogs, and social media websites. Hence, diaries now tend to take a shorter form, since they are easier and faster to generate. Diaries provide powerful methods for studying various human phenomena, such as mental health, personality processing, and physical symptoms. For this research, the recording of moments of people's lives has offered numerous opportunities and challenges. For the analysis and interpretation of these diary entries, different disciplines have been explored: psychology; sociology and computer science, which have all offered further insight into the data.

This chapter is organised as follows. Firstly, defining what *a diary* is; including its forms and development, as well as *diary methods*, their uses, collection and analysis. This is followed by a review of relevant studies from the literature on *diary features*, which in the context of this work are *activities, people, places* and common *emotions.* It then goes on to survey the *text mining* techniques that will be used with respect to this research. Finally, *data collection* methods and analysis are discussed at the end of this chapter.

## 2.2 A diary

A diary is a collection of personal information accumulated repeatedly over time. In diaries, people report *what* they do in their life, *how* they spend their time, *who* they spend their time with, *where* they spend their time, possibly their emotions, and much more. A typical traditional diary keeps daily summaries which have been recorded by the diarist in written form. It involves unstructured chronological records of a person's life events from their perspective [41]. Nowadays, many applications have emerged, such as *Day One*[1], which is a kind of micro-journelling application that allows users to record and store their everyday moments, including details of their life, memories, and photos. Another application is *Swarm*[2], which allows users to share their life logs and locations with their friends. As diary writing is a vital habit for many people, for several years diaries have been an important resource of data that have valuable applications in multiple fields: social science; psychology; higher education; health care, and many others. This research method aims to capture participants' daily experience and is commonly referred to as *Diary Methods*. The method involves intensive, repeated self-reports on events, emotions, pains, reflections or interactions near the time of their occurrence [42], and permits the examination of events in their natural context and reduces the errors that can arise from retrospection [43]. They can be used for studying the "particulars of life", obtaining person-level information and studying within-person changes, find relations between the different life events and how they affect one's mental health [44], to investigate various human phenomena [29], and to study family processes[30].
There are three main types of diary methods used in the literature:

- *Signal-contingent*: where participants receive prompts at random or fixed times to record their response. This design is suitable for measuring constantly changing variables, such as depressive mood; however, is not ideal for long studies as participants may lose interest and cease recording.

[1] https://dayoneapp.com
[2] https://www.swarmapp.com

- *Interval-contingent*: where participants are asked to record their experience at predefined and fixed intervals. These methods are suitable for routine activities or easily remembered things; however, as they are focused on the context, results cannot be generalised.
- *Event-contingent*: where participants report their data each time an event occurs, and they select when and what to record. This design is the most reliable for special events (e.g. social interactions), and not necessarily frequent, and studies can be sustained for a long time [30, 44, 45].

The data collection method in this thesis is event-contingent, where participants were asked to record in their diary as a short entry of what they are doing, including any interactions with other people and locations if applicable.

Research techniques that are similar to diary methods are referred to by various names, depending on which field the researcher is working within. For example, *experience sampling methods* [46], are similar to diary methods, but with some differentiation through the level of control the participants have when reporting. In experience sampling methods, the participants respond to surveys that are randomly or continuously generated by the system, in diary methods participants are generally free to decide when and what to record [47]. Also, a similar method is *ecological momentary assessment* [48] which combines subjective experience and related elements from the environment. Overall, a review of the literature reveals that there are no sharp distinctions between these methods in different fields; their names differ primarily based on their users and their strategy used in applying these methods.

With the emergence of new technology and the Internet, diaries took various forms. These have impacted on the way diaries are written, what is written [49] and the personal preferences regarding exactly what they want to expose [50]. Some examples are online journals and weblogs, which are modern forms of self-reported diaries intended for a broader audience when compared with the traditional diary that was considered a very personal, private record [51]. Online diaries/journals are parts of personal home pages and used by their owners for recording daily events or emotions [52]. Blogs are small personal websites that are owned and maintained by a single person and include regular updates about the author [52, 53]. Micro-blogs, such as Twitter and Facebook status, allow users to share their life events and express themselves [28].

## 2.3 Diary features

As an attempt to understand humans by reading their diaries, diary features are extracted as measurable properties of an individual. Gershuny [54] defined diaries as continuous sequential logs of events; each event is characterised by one or more descriptive fields, such as activity, location, co-presence and subjective/emotional response. Another research by Rauthmann et al. [26] provided a working model where situation cues are defined as physical or objective elements which form the environment, including a) events and activities, b) persons and interactions, and c) places. They can be objectively measured and quantified; however, these alone do not have meaning. The situation cues are processed by attaching meaning to them from personal aspects, such as associated emotions. Accordingly, in this work, a personal diary is defined by its components: external environment elements (activities, persons, places) and an internal environment element (emotions) as shown in Figure 2.1

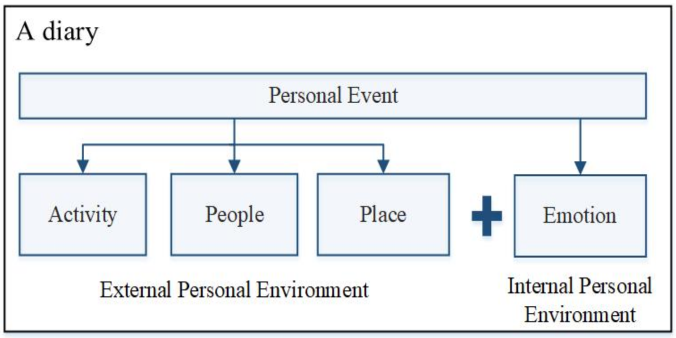


Figure 2.1: Diary features

### 2.3.1 Personal activities

A key element that diary entries often include is a personal activity that the recorder of the diary is carrying out during the day. The ability to identify daily human activities is fundamental in behavioural science, and it is an important task in computing. One common way to collect such information is by diary methods where subjects are asked to record their daily activities over a period of time. As a result, a collection of large databases of unstructured textual data is generated. Searching through this data is a time-consuming task. Therefore, automating the process of extraction is extremely valuable as it decreases the workload on the researchers and allows them to focus their attention on the analysis of the results [55]. Nowadays, data mining and advanced technology have introduced new possibilities for data exploration. The following section presents some key studies on automatic extraction and categorisation of activities from different forms of textual data and it highlights the relationship between daily activities and emotions.

#### 2.3.1.1 Activity identification and categorisation

Studies on activity identification can be viewed based on three interdependent factors: purpose; methodology and activity taxonomy. Each has a purpose and uses an appropriate method to achieve the aim and accordingly produces different activity categories. For example, studies in psychology to determine the work-life-balance may categorise activities into work and leisure, sub-categorising leisure time into social and physical and low effort activities [56]. In an attempt to build a digital diary system that helps users to keep track of their daily activities [50], personal activities were classified into nine groups based on questionnaires and users' studies: meals; drink; hygiene; work; spare-time; household; social; transport and miscellaneous.

Recent researchers have attempted to analyse real-life activities from textual data to serve different purposes. For example, Maghrebi et al. [57] extracted people's daily activities from tweets along with spatial information to utilise intelligent transportation services. Their method was to extract keywords from the text, populate a dictionary and check the occurrence of keywords to group activities accordingly. The activity type was assigned to each group of words, and then tweets were tagged accordingly. Activity groups included: eating; entertainment; shopping; social; work; recreation and home. Another study that analysed tweets to extract travel attributes like trip purpose [58], used Latent Dirichlet allocation LDA, a hierarchical Bayesian-based approach to find similarities between categorical values and explore the topics in texts. Categories found were shopping, work, entertainment, study, eating and socialising. In a study by Tandon et al. [59], they provided a vast knowledge base of human activities mined from Hollywood textual narratives. Their methodology has three main steps: semantic parsing; graph interface and taxonomy construction with the purpose of tagging videos and images. The resulting collection consisted of around one million human activities e.g., climb a mountain, kill a bird, etc. and included participating agents, information about activity sequences, spatio-temporal data and links to visual contents.

These studies have shown that the success of activity prediction depends on the different purposes to extract and categorise activities from the text using suitable mining techniques, which leads to the conclusion that it is hard to find the perfect classification for everyday activities [50]. This research will focus on the high-level framework rather than tuning an individual technique.

#### 2.3.1.2 Activities and emotions

Many psychological studies have highlighted daily activities as an important factor that affects the human emotional state and well-being. Generally speaking, the literature suggests that there is a strong correlation between emotions and everyday activities. For example, a strong relationship and constant association have been found between social activities (e.g., family gathering, parties, etc.) and positive affect [60]. Also, leisure-time and physical activities were found to be highly related to positive affect [61]. There are a large number of published studies on the impact of work on daily emotions; some studies related work as a daily activity to the increase of feelings of happiness [62], other studies relate work to a negative affect for different reasons like incivility [63], or the involvement in undesirable work events [62].

In the field of computing, researchers have focused on predicting emotions based on daily activities. For example, a study by Roshanae et al. [64] used activities in association with many other features (i.e. user engagement, gender, number of followers, etc.) to predict authors' emotions from tweets. They used Linguistic Inquiry and Word Count (LIWIC), which has personal concern categories that include 19 various activities, such as: sports; TV; eat/drink; music; money; religion; leisure; home etc. These features were used to train an SVM classifier to differentiate between positive, negative, and neutral emotions. Other researchers [65] have categorised daily activities into eight groups, using a bag of words approach with a recurrent neural network in their attempt to predict the emotions of participants in the online treatment of depression. They considered: work; recreation; necessities; exercise; sleep; sickness; rumination and social and investigated the corresponding positive and negative effects on an individual's daily emotion level. Another study on inferring the average daily emotion level based on users' smartphone logs is by LiKamWa et al. [66], using social interaction and routine activities as prediction features and an applied regression algorithm. Social interaction is represented by phone calls, text messages, and emails signals, while routine activities included browser history, application usage and location history. Yet another study extracted daily activities from life logs (i.e. physical activities, biometrics, sleep quality, diet, etc.) and used them as features to predict emotions [67]. Physical activities in their context were: time spent at home; time spent at work and time spent communicating. The state of emotions used was based on Thayer's two-dimensional model with emotional coordinates: happy; content; anxious and depressed.

Whether from psychology or computer science, the studies mentioned above have investigated the effect of activities on emotion, with varying definitions for human daily activities. Some extracted activities directly from text sources such as tweets or online focus group platforms, while others analysed personal logs (e.g.phone calls or biometrics) as indicators and used them as features to predict emotion. Researchers used different emotion models, including the common ones such as Ekman, positive and negative or even measuring the emotion level but rarely in comparison. To address this, more investigation of the emotion model is provided in section 2.3.4. This research aims to understand if there is a difference in accuracy with different types of positive and negative emotions.

### 2.3.2 People

Those people with whom a person spends time with will have an impact on one's emotions and well-being and this impact has been the focus of much research. In a study by Lin et al. [68] which used phone activities (e.g. calls and messages) as predictors of a state of wellness, it was found that the people in someone's life (more specifically the social circle of an individual) were a better predictor of self-reported happiness, stress, and well-being level than data derived from wearables like physical activities, heart rate, and sleep. This was supported by Bruce et al. [69], who emphasised in their study that in-person social interaction and connecting with friends and family on a daily basis reduces emotional stress. Other studies such as that carried out by Von [11] related emotion variables to the individual's situation, specifically noting that people experience a better mood in situations that are related to leisure such as being in good company (e.g. with a friend). One study on "very happy" people, [70] found that although no sufficient variables for measuring happiness, social relationships, especially good ones were necessary to feel happy, i.e. people who are highly social and have social interactions or romantic relationships are happier and experience positive feelings most of the time and occasional negative moods. Zhang et al. [71] tested whether the effect of the relationships with friends/parents are related to a depressive mood and found in their intra-individual analysis that emotions fluctuated with the changes in social relationships.

### 2.3.3 Places

A common element provided in diary entries is a mention of places. There are many studies which indicate that people's emotions are linked to their locations. For many people, particular places have different emotional effects. For example, Korpela et al. [72] classified places into favourite and unpleasant places and investigated the emotions experienced by people; one of their findings is that relaxation is an emotion that most people experience in their favourite places. Places that individuals visit also affect emotions in different ways and may usually be associated with some other factors such as people and activities. In a study conducted by Sandstorm et al. [27] in which they examined how people feel in different locations, they used a smartphone application that asked the participants "where are you right now?". They allowed them to choose from a list of places including home, work, restaurant/pub/cafe, family/friend's house, in transit, and other. In addition to these self-reported locations, they used the location sensed by the smartphone sensors with the question "how do you feel" and gave them the option to tap their emotion on a two-dimensional affect grid, where the x-axis denoted valence from negative to positive and the y-axis denoted arousal from sleepy to alert. A partial result from this study was that people reported positive emotions in social places more frequently than they did whilst at home and work. Other studies investigated the frequency of visiting natural favourite places, negative emotion regulation and perceived stress as well as the positive feeling of relaxation [73]. Finally, it has also been demonstrated that emotions vary based on the current situation, specifically noting that people experience a better mood in situations that are related to leisure (e.g. at home) [11].

### 2.3.4 Emotions

Emotion is an affective state that has been widely studied in psychology and has also attracted the attention of researchers in computer science. This section discusses the emotion classification schemes that have been founded in psychology based on human experiments (rather than NLP approaches). Various models exist that describe and measure emotions, including the categorical organisation of emotions, dimensional organisation and hierarchical organisation.

#### 2.3.4.1 Categorical emotion models

Categorical models are built on the belief that humans have a set of basic emotions that are discrete, measurable, and universal among people [74]. One of the most well known examples is *Ekman's six basic emotions*: happiness; sadness; anger; disgust; fear and surprise. His model was developed based on facial expressions where participants were asked about their emotions in certain situations and photos of them were taken and assigned with emotions by observers. These six emotions were found by subsequent investigators to be distinguishable across different cultures [74–76]. Another well-known model is Plutchik's *wheel of emotion* [77] (Figure 2.2) with eight basic emotions: joy; trust; fear; surprise; sadness; disgust; anger and anticipation. Emotions were arranged in a way that shows which ones are opposite (e.g. joy is the opposite of sadness) and how they can be combined to form complex emotions. The intensity of an emotion increases moving towards the centre and decreases moving outwards. There are many other categorical emotions frameworks in psychology, such as those proposed by Tomkins [78] and Izard [79]

It should be noted here that categorical emotions models do not adequately cover the complete variety of emotions expressed by humans because of their size limit. This may enforce the subjects to choose an emotion that may not precisely express their true emotion, but one which is closest to what they feel [80].

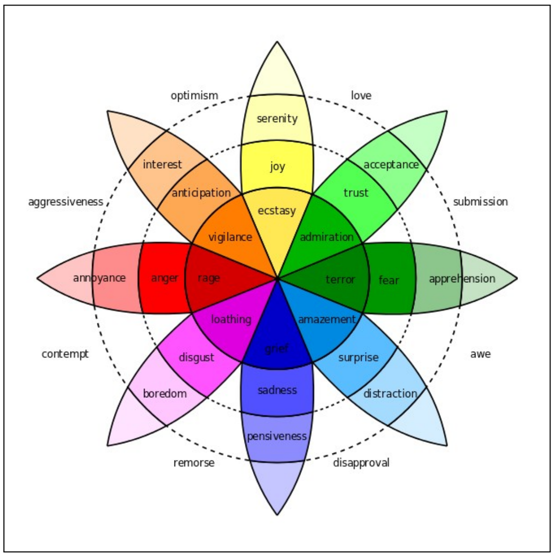


Figure 2.2: Plutchik's wheel of emotion

#### 2.3.4.2 Dimensional emotion model

Unlike the categorical emotions models that define emotions as discrete categories, dimensional models describe emotions as a set of dimensions by defining their positions in a multi-dimensional space [81]. For example, Russell's *Circumplex model of affect* [82] (Figure 2.3) suggests that emotions are distributed in a two-dimensional space containing valance and arousal. The horizontal dimension represents pleasure, while the vertical access presents activeness or how likely one is to take action. To build this model, subjects were given a sorting task and were asked to categorise words and phrases that people used to describe their emotions into one of eight categories, followed by a second task that was to place them into a circular order. Other examples are Watson and Telogen's Circumplex theory of affect [83] (Figure 2.4) and Scherer affect model [84] (Figure 2.5).

#### 2.3.4.3 Hierarchical emotion model

Emotions are organised in hierarchies to capture the property of emotions that some are sub-types of others. In most hierarchical models, positive and negative classes represent the most super-ordinate classes. The next level is considered more basic emotions (e.g., happiness, sadness, anger etc.) which are further subdivided and so on

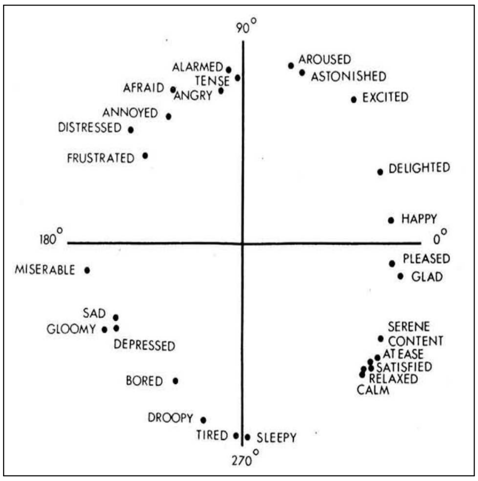


Figure 2.3: Russel's Circumplex model of affect

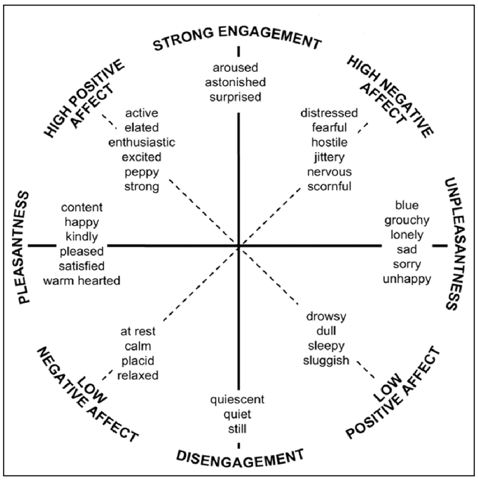


Figure 2.4: Watson and Tellegen's Circumplex theory of affect

[85]. For example, *Parrot's* collection of emotions organises more than one hundred emotions in a tree-structured list at three levels: primary, secondary, and tertiary [86]. Another example is *WordNet affect* [87], a lexical resource for affective terms, that can be used to extract emotions from text. It was derived from the WordNet database, where words with affective meaning were extracted and manually assigned with affect-labels and organised into a hierarchy (see Figure 2.6 for a fragment of WordNet affect hierarchy, taken from [88]).

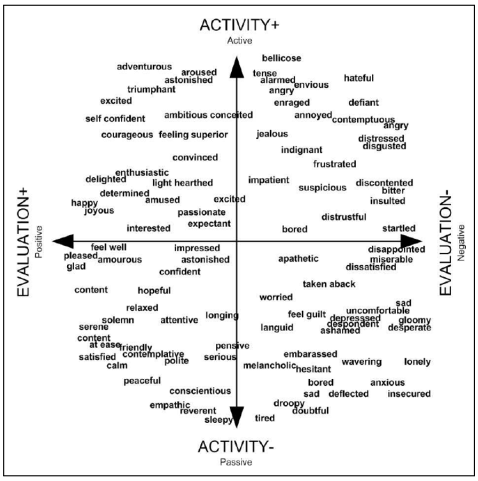


Figure 2.5: Scherer's affect model

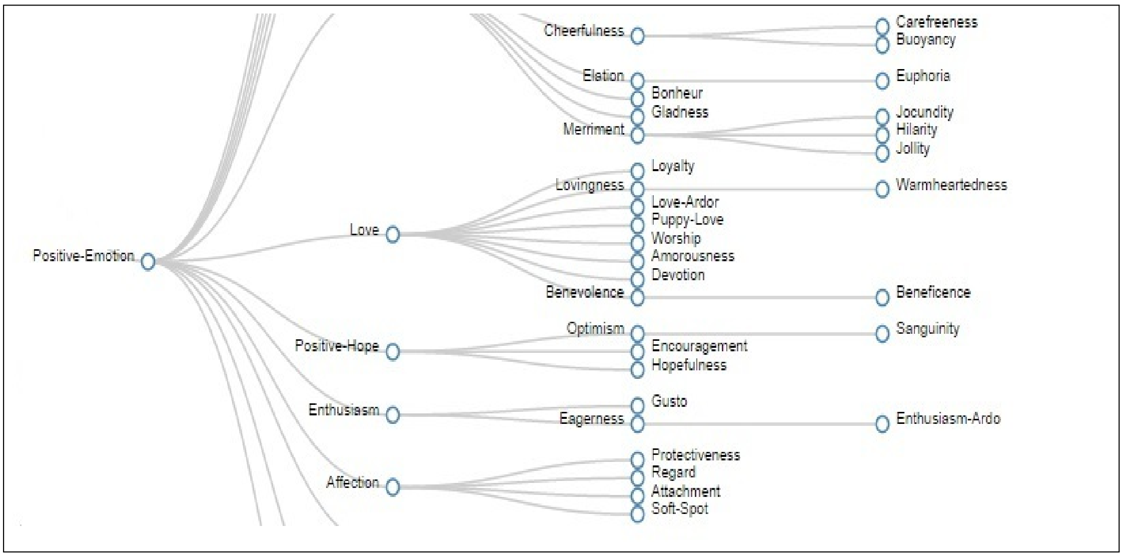


Figure 2.6: A fragment of WordNet affect hierarchy

#### 2.3.4.4 Further examination of the models

Although Plutchik's wheel of emotion (Figure 2.2) is considered as a basic categorical model, emotions are represented in a dimensional structure as they gain their meaning from their position, which illustrates various relationships across emotions, as explained earlier, Russell [82] who proposed the Circumplex model stated that Plutchick's model is similar but not identical to his model. Moreover, although Ekman treated each emotion as a separate dimension, he stated after forty years of research that pleasant/unpleasant and active/passive dimensional models of emotions are sufficient to show the differences between emotions [89]. Accordingly, there is no universal agreement on which model is the best in structure, organisation or characterisation.

The two models: Ekman's and Circumplex models, are employed and considered in this research since they are simple well-studied/used and quick to administrate. Ekman's model maybe oversimplified, representing positive emotions in only one affect word 'happy', while the Circumplex model provides a broader and more balanced range of positive and negative emotional states.

The literature suggests that there is a strong correlation between emotion and the three separate features of activities, people and places. Different psychological studies have highlighted how each is an important factor with a strong impact on emotional state and well-being. Accordingly, many approaches were conducted to predict emotion using these factors, based on different goals and the available tools.

Although this relationship is well-studied in the field of psychology, it is only recently with the evolution of technology that predicting emotion has gained interest for researchers in computer science. However, all studies show that emotion prediction remains a difficult problem, where it is hard to get high accuracy, and combining features can improve success.

While much research in the computing field has investigated emotion prediction from activities identified from sensors, images and video references, very little research has investigated activities extracted from text. Also, research has not looked at combining activities with other associated factors from the relevant event. The research in this thesis proposes a novel investigation of using a combination of these three factors (activities, people and places) as digital features to predict emotion from short diary texts.

Given the promising potential to the application of event extraction- represented by activities, people, and places- and associated emotions, it is worthwhile to investigate the appropriate text mining techniques for this purpose.

## 2.4 Text mining

Text mining is a knowledge discovery process of analysing natural language to extract meaningful information for a particular purpose [90]. This field has received a great deal of attention over recent years due to the tremendous amount of digital text data that has been available, which are generated in a variety of forms such as social networks, blogs, news, etc. Text mining is used in many different fields, such as information retrieval, natural language processing, text analysis and database technology [91]. It has many applications in different domains like marketing [92], media [93], telecommunication [94], banking [95] and health [96].

Text mining deals with the machine support analysis of the text and it is frequently found in the literature to connect machine learning algorithms with techniques from information extraction [97]. In this research, both information extraction and machine learning algorithms were utilised to categorise daily activities, extract people and places and predict emotions.

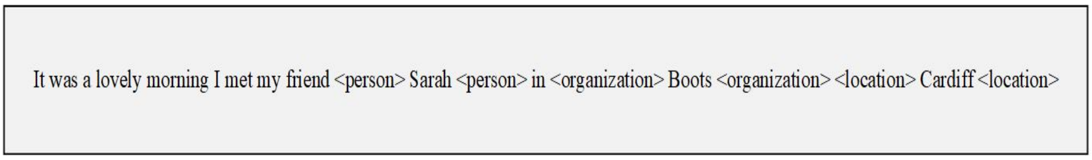
### 2.4.1 Information Extraction

Information extraction is the process of automatically extracting information from unstructured data. It is concerned with filtering information that is explicitly or implicitly stated in a large volume of text [98]. Information extraction has been widely used in many areas, for example in natural language processing and web mining and it has many applications in different domains, such as in business and social media (see [99]). Two types of information extraction methods are discussed here, *named entity recognition* and *lexical based methods.*

#### 2.4.1.1 Named Entity Recognition (NER)

NER is a sub-task of information extraction for identifying and classifying specific types of information elements into predefined groups, such as people, organisations, location, monetary values, percentages and other miscellaneous terms [100, 101], it typically applies machine learning with: word-level features (e.g., case, punctuation, numerical values, special characters, part of speech); list-lookup features (e.g., general dictionaries, last name, celebrity, country, etc.) together with a rule system over the features. For example, capitalised words are candidate entities, or a length greater that 3 words is a candidate entity of an organisation [102]. Named entity recognition has many different techniques, types and applications (see [103] for a survey).

Stanford Named Entity Recognizer is a Java implementation of a named entity recogniser. It is a part of StanfordCoreNLP, an integrated suite of natural language processing tools for English. In this research, using nltk and pycorenlp in Python, the three classes tagger is utilised for detecting Person, Organisation, and Location entities. It takes un-annotated blocks of text and produces annotated text with labelled entities. For example:



It was a lovely morning I met my friend <person> Sarah <person> in <organization> Boots <organization> <location> Cardiff <location>

#### 2.4.1.2 Lexicon based methods

WordNet is a lexical database of English language which includes nouns, verbs, adjectives, and adverbs. It is commonly used as a resource for many applications in natural language processing, it groups words into synsets, or sets of synonyms. It contains semantic relations between these synonyms, as well as short and general definitions [104, 105]. It has many different applications in text mining, including text categorisation [106], and text clustering [107].

### 2.4.2 Text encoding and machine learning

Machine learning is the area of artificial intelligence involved with automatic detection, using algorithms and statistical models, of meaningful patterns in data to be able to predict future data. Thus, it is about constructing systems that improve through experience [37, 108]. Machine learning approaches are usually categorised into supervised, unsupervised, and semi-supervised. Text data is a type of unstructured information, which is simple and easily perceived by humans but hard to process when it becomes big. Applying machine learning techniques represents a powerful way of improving understanding. However, mining large documents of text is a complex process and needs a data structure to facilitate further analysis. A common way to represent text documents and determine their relatedness is the **Vector Space Model** proposed by Salton [109] , where the documents are represented as numeric feature vectors such that the distance between them represents the degree of similarity [110]. For text classification, each document is represented by an array of words. Two commonly used text representation models are Bag of Words and Word embedding.

**Bag-Of-Words (BOW)** is a traditional feature extraction method, where the text is represented as an unordered group of words, i.e., a set of words (vocabularies). It is considered a high dimensional space representation (thousands of dimensions). There are many ways to use these vocabularies, for example, some represented the document based on the term presence (known Boolean or One hot or Binary vector), assign a value of 1 if the word appears in the text and 0 if it does not appear, others assign weights using TF/IDF (Term Frequency-Inverse Document Frequency), where a term takes its weight (importance) by calculating how often it occurs in a document and how frequently it occurs in the entire document collection [110].

Another model for text representation is **Word Embedding**, where individual words are represented as real-valued vectors in a pre-defined vector space. Compared to BOW, this model is considered a low dimensional space representation (e.g., tens or hundreds of dimensions). The word dimensions in the vector space model have no specific meaning but represent a word through the contexts where it has been found. Accordingly, it represents the distance between words by placing semantically similar words close together in the embedding space [111, 112].

#### 2.4.2.1 Supervised learning approaches

Supervised learning methods for text classification aim to assign natural language documents to predefined categories based on their content's features [113], where a set of rules are learned from a set of annotated training examples. The training set consists of text instances and labels of the category to which they belong. Usually, text instances are represented as n-dimensional feature vectors. Labels, usually, are assigned to instances by human annotators to create the ground truth; the data is then split into training and test. The class labels of the unseen test instances are then predicted based on learning the training set [37, 91].

Text classification has been used in many diverse applications ranging from large documents to short posts. Examples include sentiment analysis to classify opinions [114] and movie reviews [115], email spam filtering [116], and Twitter news classification [117].

Due to the rise in popularity of Twitter, short text classification has been extensively studied during recent years. For example, a study on tweet classification by Sriram et al. [118] used features extracted from an author's profile in addition to textual features to classify the text into a predefined set of generic classes such as news, events, opinions, deals, or private messages. They extracted one nominal feature (author) and seven binary features, including the presence of slang, time-event phrases, emphasis on words, opinion words, currency and percentages signs. They showed that using this small set of discriminative features to train a Naive Bayes classifier performed significantly better than BOW against this set of generic classes. In another study conducted by Lee et al. [119], tweets were classified into eighteen categories, including: sports; politics; technology; fashion; food and drink; health; music etc. They proposed two data models: a text-based model and a network-based model. In the text-based model, the data was labelled using human annotators. Then, after pre-processing (e.g. removing hyperlinks), the documents went through a string-to-words vector kernel to filter out common words using TF-IDF and extract the language features. In the network-based model, they used Twitter-specific social network information such as the tweet's time and number of tweets related to a topic, and followers. They compared the performance of different machine learning algorithms for classification, including: Naive Bayes; Naive Bayes Multinomial; Decision Tree; Support Vector Machine; K-nearest neighbour and ZeroR. Their key findings were that network-based performed better than a text-based approach and that Naive Bayes Multinomial returned good results with the text-based model.

Examples for algorithms that have been proven to work well for many classification tasks, including events extraction from personal generated data [120] and mining short texts, [121] are: Support Vector Machine (SVM); Naive Bayes (NB) and Decision Tree (DT).

**SVM** is a supervised discriminative algorithm that is characterised by producing a hyper-plane that separates the data into different classes and derives the widest channel between the classes [122]. It can be used for linear classification as well as non-linear classification using kernel function [123]. The main advantage of this algorithm is that it is very resilient to over-fitting.

**NB** is a probabilistic classifier that measures the probability that a document or word is a member of a category. This model deals with each word independently. It has been proven to be a powerful approach for many reasons: it is simple, easy to implement, fast to compute and efficient when compared with non-Naive Bayes approaches because of the words' independence [124].

**DT** is a classifier in the form of a structured tree, as its name implies. Each node can be either a leaf or a decision node. A leaf shows the class label, while the decision node implies a test to classify the data. Internal nodes are labelled by attributes (word occurrence in our case) and branches departing from them are labelled based on the weight that an attribute has in the text document. It categorises the data by recursively testing the weights of the attributes and labelling internal nodes until a leaf is reached. It can easily be tracked from the input (root) to the output (leaf) and needs less training time since it does not need any parameter settings [122].

Supervised approaches can achieve high accuracy. However, it is costly and time-consuming, especially for labelling big data sets [123], and does not generalise well to different domains. Moreover, text annotation may be subjective. Consequently, semi-supervised methods are proposed to start classification with a small sample to labelled training, in addition to a large unlabeled data.

#### 2.4.2.2 Unsupervised learning approaches

Unsupervised learning methods aim to find hidden patterns from unlabelled data where no training phase or manual effort is needed [125]. Text clustering is an unsupervised learning method. It statistically segments a collection of text documents into partitions based on some distance measure in order to identify natural groups where documents in the same cluster are more similar to each other and between clusters that are dissimilar [37, 125]. Unlike supervised learning in which categories are known before classification, in clustering, there are no predefined classes and, therefore, there is no need to label data in advance. Thus, it is flexible and has high automation handling ability [126]. This makes clustering exploratory in nature, as it aims to find structures in the data [125].

Text clustering has a wide range of applications, such as documents classification and information organisation [127]. Word clustering has three advantages: it returns useful semantic words groupings; increases the classification accuracy and reduces the size of the classification model [126].

The literature has shown that word clustering is a powerful alternative to feature selection when applied to text classification [128]. For example, Dai et al. [129] proposed a word-embedding based clustering method to classify tweets, where each tweet was represented by a few vectors of their words. Accordingly, they were divided into clusters of similar words to identify whether each cluster was related to 'flu' or not. They used Google's pre-trained vector set trained by Google's news data set for their research purpose. As this method preferred using the semantic meaning (i.e., word embedding) over BOW, another study proposed a combination between BOW and word embedding in an attempt to gain the advantages of both and overcome their weakness. Kin et al. [130], proposed a bag-of-concept method, where concepts are created by clustering words based on their low-dimensional vectors generated from a basic neural network model, (word2vec), and used the frequency of these concepts to represent the document vectors. One of the data sets they used was the Reuter data set which has articles labelled into eight different classes, including: entertainment; sports; technology; politics; business; market; world and health.

#### 2.4.2.3 Semi-supervised learning approaches

A semi-supervised learning approach is a hybrid sitting between supervised and unsupervised learning methods; it addresses the problem of having insufficient labelled data by using a large amount of labelled data with a small amount of unlabelled data for building a good classifier [131]. Similar to *bootstrapping*, where a small number of instances are used as seeds or patterns for bootstrap learning, it is an iterative process where these seeds are used with a large corpus to extract patterns to enrich the learning process and used to extract new instances [132]. Although this approach needs less labelled-data, and thus carries a lower cost, it is sensitive to the original set of seeds and requires that seeds are present for each class, otherwise the resulting patterns suffer from low precision and semantic drift [131]. A good example is the approach in [121], where researchers proposed a semi-supervised method in combination with SVM to classify short text into one of five classes: politics; economy; education; entertainment or science. They applied iterative learning and labelling the unlabelled samples by calculating the similarity between samples until all were labelled entirely. They showed that this method improved the classification of short texts.

### 2.4.3 Transfer learning

*Transfer learning* is a machine learning method concerned with improving a classifier from one domain by transferring information from a related domain. It aims to create a high-performance classifier trained with easily obtained data from a different source domain. Its objective is to design a system which can leverage experience from previous tasks when a new task is drawn from a different population than the old one and to improve performance in a new task which has not been previously encountered [133–135]. In contrast to traditional machine learning, which predicts labels of future data using statistical models that are trained on labelled or unlabelled previous data available for the problem of interest (training), transfer learning allows the domain, task or distribution to be different between the training and testing data [136]. When training data is expensive, insufficient, or difficult to collect the need for transfer learning occurs, which decreases the time spent learning new tasks and requires less human intervention [134]. Transfer learning can also be referred to as knowledge transfer [135], learning from auxiliary data [137], or domain adaptation [138]. The common scheme is to use data from a related source domain to train or improve a target learner. While there is a general acceptance and comprehension of the "transfer learning" concept, usage and application varies greatly. One example is the domain adaptation for sentiment classification proposed by Blitzer et al. [139], who selected Amazon reviews for different products: books; DVDs; electronics and kitchen appliances. In the task of automatically classifying a review for a product (e.g. a camera) as positive or negative, they used a classification model that was trained for source products to classify other target products, for example, they looked at the adaptation from book reviews to reviews of kitchen appliances. Another study on system recommendations by Morino et al. [136], proposed a transfer learning technique to extract knowledge from multiple dense domains containing rich data (e.g. movies, music) to train and generate recommendations for a target domain (e.g. games). In their study, Jin et al. [137] classified short text represented in tweets and short online advertisements in the predefined topic using auxiliary long text of the content of the associated URLs for tweets and randomly crawled e-commerce web pages. Also, with web documents classification, for example, in a study carried out by Dai et al. [140], newly created or outdated web documents were classified into several predefined categories learned from using previous manually labelled web documents.

Transfer learning can be supervised [138],unsupervised, as with self-taught clustering [141], or semi-supervised [140], heterogeneous domain adaptation [142], or homogeneous domain adaptation [143]. In this research, knowledge is transferred across different domains in a semi-supervised learning process to classify diary entries into pre-defined categories of daily activities.

To summarise, recent research such as that on transfer learning (Section 2.4.3), which introduced the concept of using an external domain that is rich with similar data for training, have overcome the challenge of a lack of training data. Moreover, APIs now make it easy to collect data, and many researchers share their data sets online.

This section has displayed the main machine learning algorithms and discussed the main text representation methods: BoW and word embedding. However, BoW suffers from sparsity when the number of unique words increases and assumes every word to be independent while word embedding reserve proximity, which defines a semantic relationship between the words. Researches have been extensively using both methods as needed. This section discussed some research that has overcome these methods' limitations by combining them (Section 2.4.2.2). Additionally, some studies mentioned earlier preferred a small set of features to train a classifier with common supervised machine learning algorithms (Section 2.4.2.1) as they showed that the performance is significantly higher than using the linguistic features.

The methodology in the research in this thesis was motivated by the concept of transfer learning using an external domain that is rich with similar human-generated training data to build the framework of activity classification for diary data, which is very personal and hard to collect as it takes time and needs commitment from the participants. The proposed approach for activity classification here aims to integrate the advantages of BoW and word embedding methods, utilising the semantic similarity of words in the tweets vector space and activity word occurrence in diaries, while maintaining low dimensionality of word embedding while providing interpretable representation concurrently. This research was also motivated and encouraged by those studies that preferred a small set of features to train a classifier with common supervised machine learning algorithms, accordingly use a small set of features (activities, people and places) instead of/with BoW for emotion prediction to compare and improve the performance.

### 2.4.4 Emotion detection

Affective computing is concerned with creating a machine that can interpret emotion by recognising, expressing and modelling emotional information [144]. Emotion recognition aims to detect different types of humans' emotional states from various sources, including text, audio, gesture, eye gaze and facial expression. More specifically, emotion detection from text focuses on finding the relations between the input text and the emotion that prompted the author to write this text [145]. That is to say, it is a classification problem that labels instances with the relevant emotion based on the content of the text. Several techniques have been proposed for building emotion classifiers which can broadly be classified into two main categories as follows:

1. **Lexicon based methods**
   These are keywords-based approaches which predict emotions by identifying an exact, specific word from the text and matching it with a predefined term. Thus, emotion is inferred based on the related keywords found in the input text [145]. They are also considered to be dictionary-based approaches, where words are defined and organised in different databases. Examples are *WordNet affect* [87] (discussed in section 2.3.4.3) and *SentiwordNet* [81] which is derived from WordNet, where each word is associated with a numerical score that indicates whether the word is positive or negative (for example 'happy' has PosScore: 0.875 and NegScore: 0.0). Although very simple and flexible, the fact that they rely on individual words may affect the performance when the sentence structure is complex, as there may be ambiguous words and sometimes there is no explicit expression of emotions in the text. Also, it is impossible to cover all the possible emotional words due to the evolution of the language [146].

2. **Machine learning methods**
   All the previously discussed methods in Section 2.4.2 for text mining can be used to solve the problem of emotion recognition from the text. However, to employ machine learning algorithms this problem has been redefined to determining to which emotional class the input text instance belongs [147]. This section explains text-based emotion prediction problems as a classification problem, and highlights some studies that employed machine learning methods, especially for emotion recognition.

Examples of **supervised machine learning** include a study conducted by Alm et al. [148] to classify emotions in the narrative text (sentences) in children's fairy tales to be used later in assigning appropriate sound in the text-to-speech analysis. In this work, they used language features like BOW, special punctuation, sentence lengths and others to train a linear classifier in a feature space using SNoW learning architecture to classify sentences into emotional and neutral sets. They labelled the emotional data set with Ekman's basic emotions and combined them to be classified into positive and negative sentences. Bellegarda [149] carried out a study that utilised Ekman's emotional scheme for the classification of news headlines, which used latent semantic mapping with Support Vector Machine, Naive Bayes, and Decision Tree on BOW in association with WordNet affect. In their study, Hasan et al. [150] classified tweets into the distinct emotional classes they express utilising the Circumplex model, with hash-tags used as labels to train and compare between different classifiers: Support Vector Machine; K-nearest neighbour; Decision Tree and Naive Bayes. They used BOW to create the n-dimensional numerical feature vectors, in addition to emojis, punctuation features (e.g. the exclamation mark). Emotion detection was also used in the classification of film reviews, for example Kennedy et al. [151] used Support Vector Machine in their study to classify the reviews into positive, negative and neutral based on uni-gram and bi-gram language features. In a related study, movie review classification was divided into positive or negative [152] and showed that traditional topic-based categorisation performed better than Naive Bayes, Maximum Entropy and Support Vector Machine.

**Unsupervised learning methods** mainly consists of sentiment clustering methods implemented for many purposes and utilises different emotion models. For example, for emotion detection in lyrics-based songs, Hu et al. [153] proposed a fuzzy clustering method to group the lyrics in their sentences around the associated emotions from the Circumeplex emotion model. Yuan et al. [154] proposed a word clustering method to improve emotion recognition from short text, where they categorised news headlines by proposing a weighting scheme for words' clusters based on discrimination degree of clusters and word representation degree. Each line was annotated with an integer between 0 and 100 that indicate the different degrees of support for each of Ekman's six basic emotions. Others extracted and used their emotion categories,

> such as Feng et al. [155] in their work on extracting common emotions from blogs towards some events. For example, to classify the public emotions towards the sudden withdrawal of the former world champion of the 110-meter from the Olympics, they grouped bloggers' emotions into nine groups, these included: disappointment; blame; deception; support; regret; understandable; inveracious; normal and non-Olympic.

The growing importance of automatic emotion identification has coincided with the emergence and growth of social media such as Twitter, Facebook and online blogs. Personal blogs are perhaps the most relevant research area in emotion detection, where people express their thoughts, emotions, opinions and personal daily activities and routines. For example, Samonte et al. [156] combined keyword spotting and semantic analysis exploiting a word position in the sentences to strengthen prediction. Also, Shivhare et al. [157] proposed an architecture for predicting emotion from blogs based on keyword spotting in combination with the use of ontology by applying a semantic approach using some aspects like entities, attributes and relationships, that describes the concept of a domain. Other studies focused more on machine learning techniques. For example, Chau et al. [158] combined machine learning classification with rule-based classification obtained from experts to identify people with emotional distress. The proposed system achieved better performance than the a benchmark, and experts found it more efficient than traditional methods.

The common text mining methods (explained in detail earlier in this chapter) used for emotion detection in blogs and social media are lexicon-based, including keyword-based techniques and lexical affinity, learning-based and hybrid approaches [147]. However, each method has its own limitations, such as ambiguity and lack of linguistic information. Researchers have aimed for improvement by combining different methods to enhance interpretation and to overcome these limitations. Accordingly, this thesis proposes to deal with these linguistic challenges by generating numerical probability feature vectors that describe the activity, instead of using exact emotion words as features, to be combined with people and places for emotion prediction.

## 2.5 Conclusion

This chapter explored the diary and explained some of its features, including daily activities, people, places, and emotions. By highlighting research from both the fields of psychology and computer science, it has shown the effect external environmental circumstances (activities, people, and places) can have on an individual's emotion. It also provided a broad view of the current text mining techniques and highlighted subtasks, information extraction and machine learning methods that have been proven to be powerful approaches for text processing.

Drawing on this past research into human emotions in the discipline of psychology and using the well-studied techniques in the machine learning field, the present research investigates the capability of these technical methods to be applied for diary analysis and understanding. Accordingly, the work in this thesis is built by using available text mining algorithms to make the data computable, that is to express the data as mathematical analysis and processing of form, automatically extract features from diaries and infer relevant emotions.

# Chapter 3

# Activity categorisation framework

## 3.1 Introduction

Many of the everyday events and activities which people experience during the course of their lifetime have an impact on their emotions and well-being. Various activities, from cleaning the house to working on a project, or having lunch with a friend, affect their emotions in different ways. This chapter focuses on the automatic categorisation of diary entries which describe daily activities in an attempt to utilise them as features for predicting emotions. As an example, the diary entry “having lunch with mother” could be classified as: 70% ‘food/drink’ and 30% ‘social family’. However, although such an entry is easily understood by human the automatic categorisation of diary data that contains such personal activities can be challenging.

This work presents a novel approach to the categorisation of personal daily activities based on transfer learning. The proposed framework utilises previously acquired knowledge as a starting point to solve the present problem, which reduces the time required to learn and the amount of effort necessary to label the data. A high-performance diary learner was trained with easily obtained, similar but not identical, data from a different domain (tweets) to classify individuals’ textual diaries into daily activities. Moreover, customised pre-processing steps are proposed to filter non-meaningful data to the domain to enable more efficient classification.

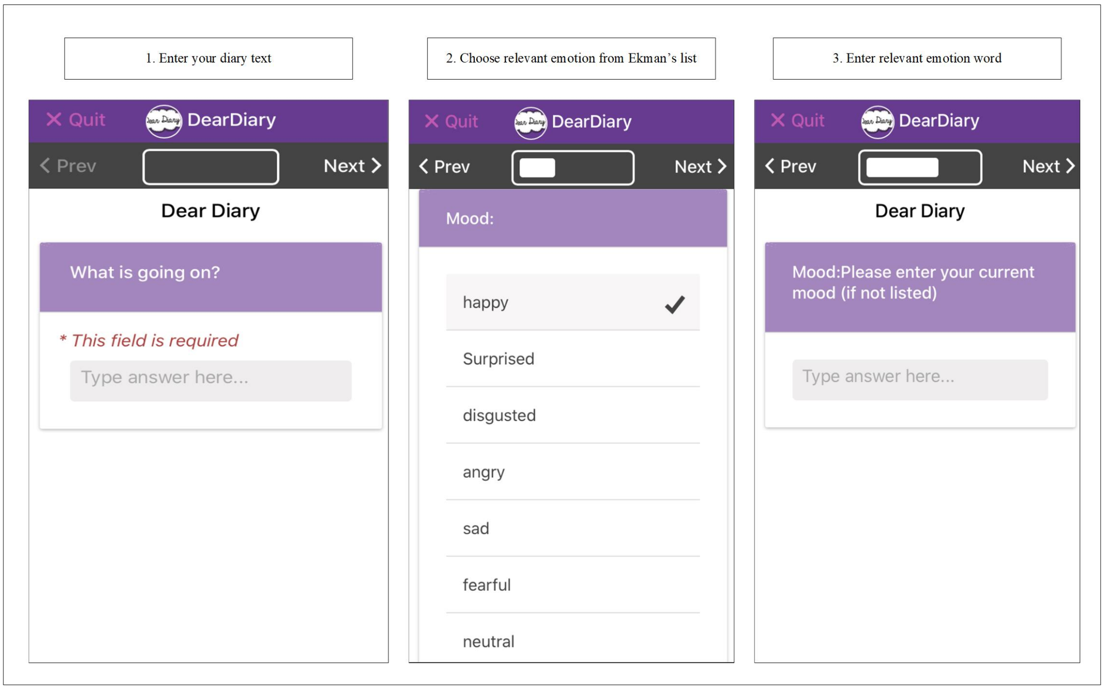
This chapter is structured as follows: Section 3.2 explains the data collection and the data set. Section 3.3 introduces personal activity groups which are used to structure daily activities. Then, Section 3.4 provides a brief overview of the proposed four-stage framework for activity categorisation, which is followed by four sections that explain each module in detail. Detailed evaluation of the framework is presented in Chapter 4.

## 3.2 Data collection

As mentioned earlier in the previous chapter, a diary may take different forms and there are various collection methods. The focus of this research is to analyse labelled diary entries; hence, a data set was collected consisting of short texts and emotions. This section introduces the data collection method untilised in this work, the data set, as well as ethical approval and consent to participate.
For the data collection, a smartphone application *Epicollect5* [1] was used, software developed at Imperial College London, it is a web-based data collection tool that was found to have all the required functionality to create a project called **Dear Diary**. Epicollect has been used for both teaching and research, for example [159, 160].
The interface provides a text box with the question *"What is going on?"* that enables the participants to record their daily diary posts along with a drop-down list of the emotions defined by Ekman together with a text box to allow the participants to choose from the list or enter an alternative word that best describes their relevant emotion. It is available for users of both the Android and iOS platforms. The smartphone has been chosen as a platform for the data collection because people carry their phones with them all the time and are always available for quick and prompt data recording. Figure 3.1 shows the data collection application interface.

The participants were asked to type their activities in text format and were encouraged to mention what they were doing at that moment, or had done during their day when they were somewhere or with someone, or even doing something alone, then to save their posts and upload them through the phone. Fifteen participants were asked to post their data at three optional times per day for two to three months. Between around 100 to 250 entries were collected from each user with an average of 161, a

[1] https://five.epicollect.net



What is going on?

* This field is required

Mood:Please enter your current mood (if not listed)

Figure 3.1: Dear Diary application

standard deviation of 56.7 and around 2500 entries in total. The response rate differed between participants due to the longitudinal nature of this study. Also, the format of the contents of the text entries varied tremendously. Some of the participants provided only keywords, while others recorded short sentences and some saved long and detailed entries, with an average of 9 words and a standard deviation of 4.85. Emotion labels vary between users; some used the given list of emotions, while others preferred to enter an emotion-word that best described their feelings associated with their entry. Table 3.1 displays some examples of the data. Further statistics are provided in Chapters 4 and 5.

The project was constructed to be a personal, private and trustworthy way for the participants to record their data. The study was approved by the ethics committee of the School of Computer Science and Informatics, Cardiff University. All participants provided written, informed consent to participate and for part of their anonymous data to be made available for research purposes.

In this thesis, the same data set is utilised and analysed for different purposes. In chapter 3 and 4 all participants' data is used for daily activities identification and categorisation. In chapter 5, with personalised emotion prediction, each participant data is treated as a stand-alone data set for training and testing. While with global, the data was used collectively for training and prediction was tested on each user's data, this is explained further in subsequent sections.

| **Diary entry** | **Emotion** |
|---|---|
| Drinking sleepy tea at home | relaxed |
| I had my dinner with a friend at GBK | **happy** |
| Having a walk with my children at the park | content |
| Sightseeing with Farah in Camden market | excited |
| I had my coffee in Starbucks with my friend Sarah | **happy** |
| I feel free today after submitting the conference paper and students projects marking | **happy** |
| Planning for a trip | confused |
| Learnt about a death of a family member | **sad** |
| I just made mistake at work and I am totally not happy I can't handle the work environment at this company | **sad** |
| Visiting a museum | Unimpressed |
| Catching up with a friend | neutral |
| Visiting the book fair with my kids | excited |
| Playing PlayStation in an open area with my friends | excited |
| At home watching a movie with my husband | enthusiastic |
| Attending football match with friends | **happy** |
| I had an interview with booking.com and I have good feeling that I will take the job | excited |
| Now I'm drinking my coffee and smoking | neutral |
| Today I went to the mall with my friends and I bought a new sunglasses | **happy** |
| Working out at gym with my friends | **happy** |
| Finally Shaving at the barbershop | **happy** |
| I cleaned my flat | neutral |
| Baking | **happy** |

Table 3.1: Diary entries examples. Ekman's emotions are shown in bold

## 3.3 Personal activity groups: definitions

In this thesis, a personal activity is considered to be any action that is carried out by a person. People mostly talk in their diaries about daily events or activities: working; shopping; visiting and so on. Since there are different activities conducted by different people, grouping them would be a convenient means by which to analyse and organise this data. However, "The perfect classification of activities of daily living is not possible" [50]. Usually, the classification of daily activities is highly dependent on the target and, accordingly, different categorisations have been proposed by different researchers [50, 56–58, 64, 65]. In line with other approaches, the data was investigated, more specifically clustered (more explanation is given in Section 3.7.1) and the clusters were manually examined; accordingly, the daily activities set out below have emerged as natural groupings.

**Work/Study group**

This group contains events that involve doing a job or studying. Typical, work-linked activities include entries such as: “meeting my supervisor in her office” or “working on a project with my team”, whilst study-linked activities are similar to “doing my assignment at the library” or “studying for my next exam”.

**Social/Family group**

This group contains actions involving a person (or people) other than the participant in a social interaction. Examples of entries for this group are: “visiting my mother” or “talked with my friend on the phone”.

**Food/drink group**

This is the group of any activities that are related to food consumption, for example “ate pasta” and “had breakfast”.

**Leisure group**

This involves activities that a participant conducts in their spare time. Examples for this category can be: “watched a great movie” or “in a play”.

**Essentials group**

All actions that need to be done by a person and do not belong to any of the above groups fall in this group like “cleaning my apartment”.

Although these are ‘natural’ groupings, not all entries will fit exactly into only one. For example, taking the entry “having lunch with my family”, could be classified as: 70% food/drink and 30% social/family.

## 3.4 Activity categorisation framework

The approach proposed in this work was developed to take a short piece of text and return a classification into activity groups. For example, a diary entry such as: “having breakfast with my family” should be classified as both food consumption and a social activity.

This section introduces the framework for activity categorisation. As a starting point, the system exploits knowledge from 'a source model' of similar data (tweets) to accelerate and ease the activity classification process of the diary entries. The proposed method can be described as a 'word-embedding based clustering-classification' that takes advantage of semi-supervised approaches to classify the data. Figure 3.2 shows the framework for the automatic categorisation of daily human activities from diary data. It consists of four main modules that are discussed in detail in the following sections.

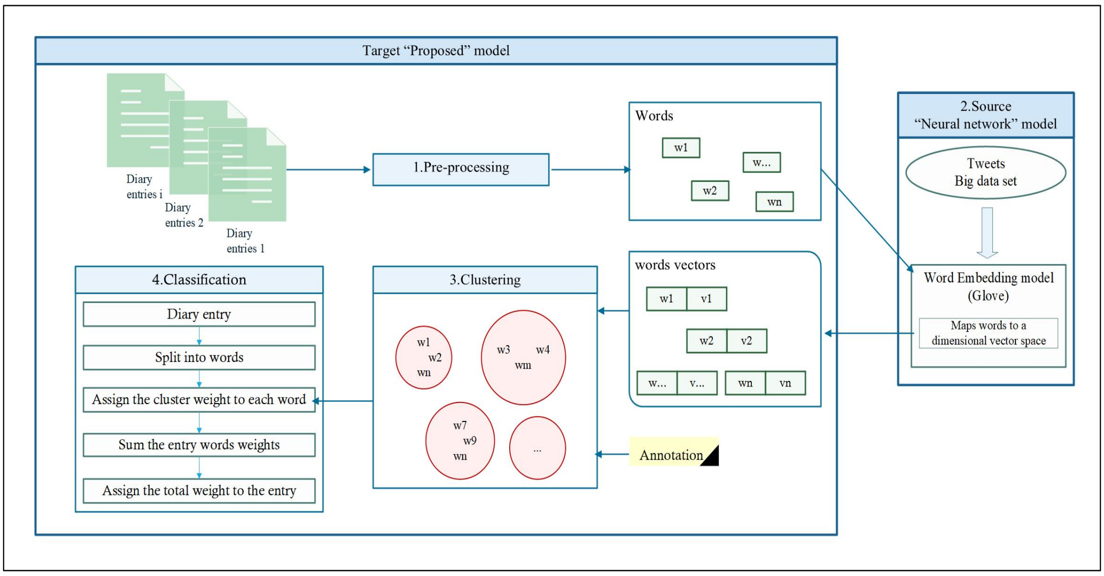


FIGURE 3.2: Activity categorisation framework

1. **Pre-processing**

   This step comes before the data analysis as entries may have some noise. Here, in addition to standard NLP pre-processing techniques, more customised steps, such as the removal of people's names, are introduced to reduce the amount of unnecessary data that consequently reduces the purity of clustering and accuracy of classification (Section 3.5).

2. **Word embedding transfer learning**

   This is the initial step for feature extraction exploiting a pre-trained word embedding model, where diary words are defined and are mapped to a vector space embedding (Section 3.6).

3. **Clustering**

   In this step, activity groups are formed from the defined diary words, weights are then assigned to the cluster based on crowd-sourced annotations (Section 3.7).

4. **Classification**

   This step aims to assign a weight for each activity type to each diary entry after breaking it down to its content and referring to the clusters for its words' weights (Section 3.8).

An example of a diary entry classification is shown in Figure 3.3.

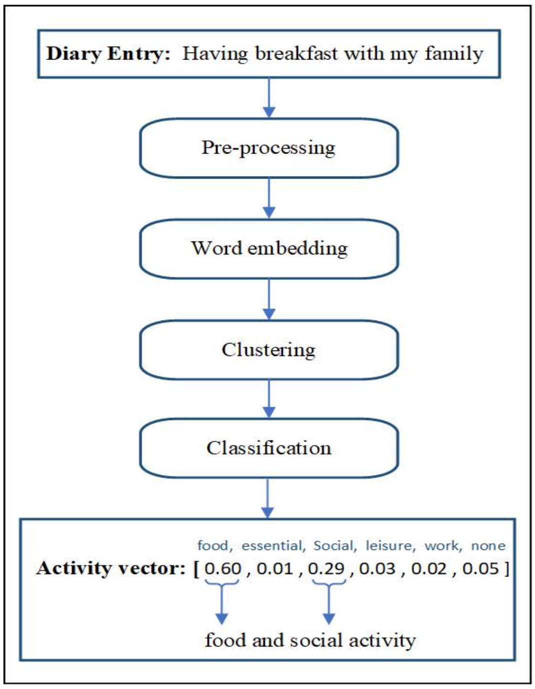
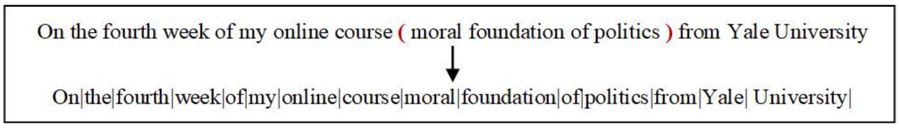

Diary Entry: Having breakfast with my family
Pre-processing
Word embedding
Clustering
Classification
food, essential, Social, leisure, work, none
Activity vector: [ 0.60 , 0.01 , 0.29 , 0.03 , 0.02 , 0.05 ]
food and social activity

FIGURE 3.3: An example for a diary entry activity classification

## 3.5 Diary entries pre-processing

Pre-processing is an important stage that usually precedes text mining. The diary entries are characterised by being informal, unstructured and noisy, where some may contain spelling mistakes and non-textual information. In this work, pre-processing was defined through an iterative process which went through refinement after running consecutive clustering experiments and observing the changes. The following are the pre-processing steps which the data went through for the final analysis:

- **Tokenisation, filtering and removal of duplicate words**. All tabs, punctuation and non-letter characters were removed, and sentences were broken into tokens (words) using the Python natural language tool kit (nltk).

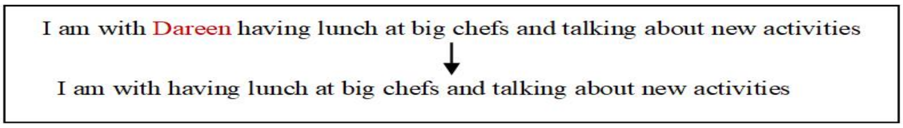


On the fourth week of my online course ( moral foundation of politics ) from Yale University

↓

On|the|fourth|week|of|my|online|course|moral|foundation|of|politics|from|Yale| University|

- **People names removal**. As people names in different participant's data do not contribute to the clustering and classification of the entries. On the contrary, initial clustering (see Section 3.7.1) distributed names through different activity clusters which reduced the purity of the clusters and accordingly may cause bias with classification (deviation or misclassification). Named Entity Recognition (NER) was used to detect people's names (ex. Sarah, Dareen, etc.) and then remove them. Python StanfordNERTagger was imported to detect names.

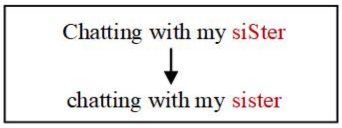


I am with Dareen having lunch at big chefs and talking about new activities

↓

I am with having lunch at big chefs and talking about new activities

- **Conversion to lower case.** To treat the words consistently, this step took place after detecting and removing all people names because the algorithm of NER checks for capitalisation of initial letters to pick the names.



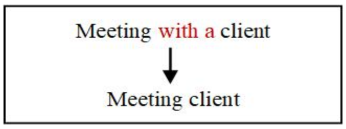
Chatting with my siSter

↓

chatting with my sister

- **Stop word removal**. These are words that are commonly used and add little meaning to the domain and the topic of clustering, such as: the; a; are etc. For this purpose, stopwords.words (English) was used to extract them.

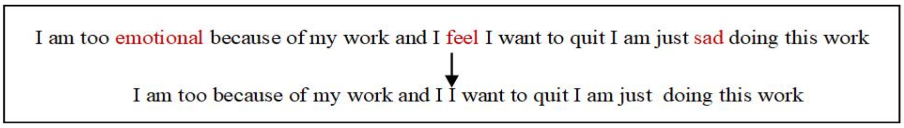
- **Emotion word removal**. Words such as hope, lovely, joy, etc. are removed using an external data source WordNet lexicographers' files:noun.feeling and verbs.emotion.



Meeting with a client

Meeting client

This step was important since initial clustering (see Section 3.7.1) accumulated such words in a single cluster; however, they added no meaning to the activity related clusters.



I am too emotional because of my work and I feel I want to quit I am just sad doing this work

I am too because of my work and I I want to quit I am just doing this work

## 3.6 Word embedding transfer learning

In order to progress further with analysing diary entries, each word should be represented in a consistent way that allows them to be compared and clustered. A foundational step in building this framework is to enable it to apply previous knowledge to the diary domain for the recognition of daily activities. Since diary data is expensive and hard to collect, the system takes advantage of a big data source of easily obtained data from Twitter. The data in this research is somehow similar to tweets in content and structure, i.e. people usually share their activities in unstructured short texts through that platform. The process of using related information from a source domain (tweets) to improve the learning function in a target domain (diary) is referred to as *transfer learning* [133, 161, 162]. Figure 3.4 demonstrates this knowledge transfer process for the representation of diary words.

In this research, an artificial neural network is utilised to represent the diary words; this architecture has proved to produce strong results in text processing [163]. It is a machine learning algorithm that has a strong capability for word embedding, where each word is represented as a continuous vector space which is generated based on the co-occurrence of words and represents the semantic relationships between words. For example, the word 'working' can be represented by the following vector:

[0.22431,-0.25557, 0.006813, -0.18593, ..., ..., 0.14495, 0.22284, -0.056669, -0.011425]

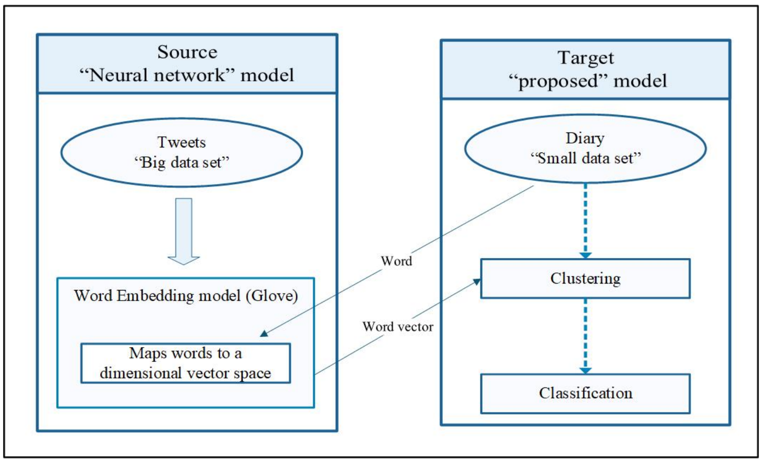


FIGURE 3.4: The transfer learning process

With this representation, words that usually occur together are closer to each other in the vector space. However, these individual dimensions in the vector space embedding have no specific meaning [164].

Although with word embedding there is a loss of interpretability, the fact that this method generates low dimensional vectors that preserve proximity made it preferable over one main document/text representation method (Bag-of-Words) that is known for its simple interpretability where a document vector is represented by its words frequency/presence. However, when the number of words increases, the BoW method suffers from high dimensionality, which causes sparsity and fails to preserve accurate proximity information.

An example of an implementation of this architecture, the Global Vectors for Word Representation "Glove" [164], is an open-source toolkit developed by Stanford, and was used here via the Python Gensim library. This model is pre-trained with 2B tweets and 1.2M vocabularies producing 200-dimension vectors for each word. Hence, each diary word was defined and assigned a 200-feature vector representing its occurrence in the Tweets corpus.

## 3.7 Clustering of diary words

Clustering is an unsupervised machine learning method designed to reveal hidden data patterns, which identifies natural groups in multidimensional data [37]. With clustering algorithms, different similarity measures are used to aggregate the data and produce different clusters. Accordingly, members of each cluster should be similar to each other in some sense and dissimilar to the members of other clusters [165]. This section discusses the data clustering process, which has three main steps, as shown in Figure 3.5:

1. **Word clustering** where words are grouped based on their assigned dimensions in the vector spaces.
2. **Word annotation** where words are manually assigned a label of the relevant activity group.
3. **Clusters-weights assignment** where all clusters are assigned weights based on the results from the previous two steps.

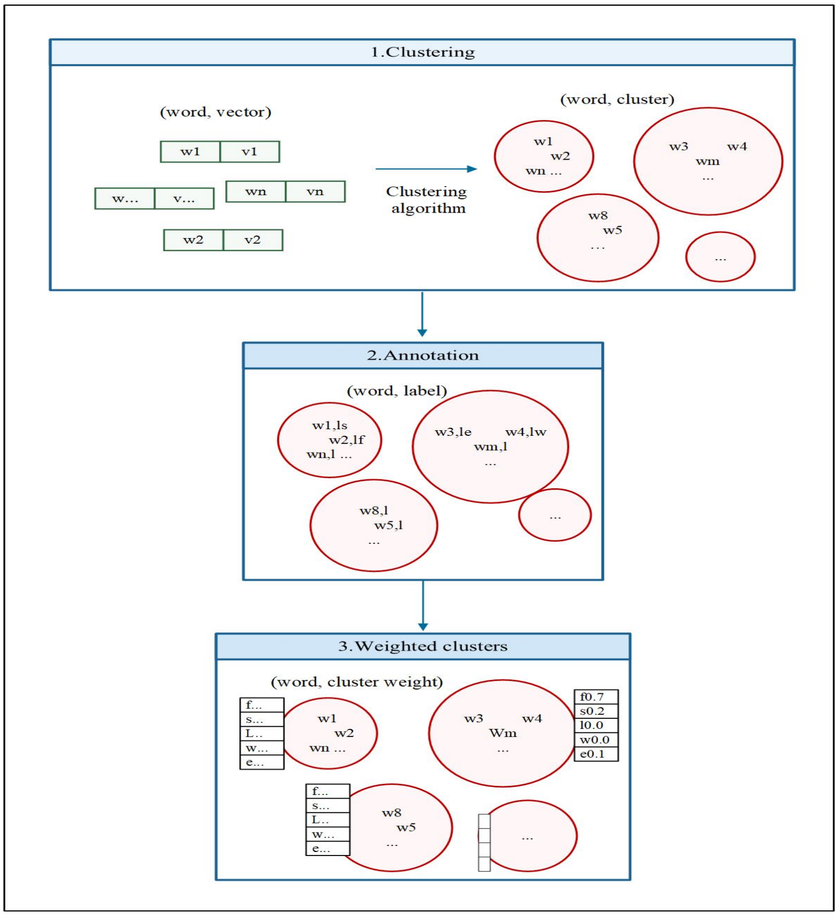


FIGURE 3.5: Word clustering framework

As shown in Section 3.6, each diary word was defined by a continuous vector that represents the meaning of a word through the context in which it has been observed in a corpus of tweets. Accordingly, if two words occur in a similar context, they most probably have a similar meaning. Therefore, in a new diary data set where new words that haven't occurred previously but can be found in the larger tweet data set, by taking its dimensional vector space and find its similarity and closeness to the cluster's centroid, the word can be assigned to the most relevant activity without re-running the clustering algorithm.

Although word embedding could be used for classification, a clustering step is essential in this framework, more specifically using a centroid-based clustering algorithm (e.g., k-means) provided an advantage that can increase efficiency, i.e., having a fixed position of a word that is close to all relevant words in a cluster is vital since new words could be added to the cluster where the nearest centroid exists based on similarity and proximity without the need of re-clustering or re-classification, of course, entropy should be considered in such case.

### 3.7.1 Word clustering

Words were clustered into groups based on their corresponding vector space dimensions, as explained in the previous section, where 200 features were used to group words. K-means was implemented for clustering as it has been shown to be a powerful and robust algorithm for text categorisation [166, 167] especially with human-generated short posts [168]. The main advantages of using this algorithm are that not only is it one of the simplest partitioning algorithms and very easy to implement [169], its linear time complexity makes it fast and, as a result, needs low computational cost when dealing with multi-dimensional data [170]. The K-means algorithm was chosen here as its a centroid-based approach which benefits the proposed framework when dealing with new or unseen data.

At the beginning, parameter setting took place. This is a challenging task in clustering where it is an unsupervised method and the machine has no previous knowledge of the data. So, this method relies on experimentation where there are multiple runs of the same algorithms with different parameter values. For k-means clustering, the main parameters to choose the right clustering validity are the number of clusters and distance function.

Automatically determining the number of clusters $k$ has been one of the most complicated issues in data clustering [125]. Different methods are used to assist with clustering numbers estimation, such as the elbow method [171] and gap statistics [172]. However, they are not guaranteed to give the optimal number of clusters that suits the domain and model as they are considered intrinsic measurement criteria [173], meaning that they are domain-independent and model-dependent. Therefore, following what is common practice, the algorithm was run with different values of $k$ then the best number of $k$ was chosen based on the model, the domain and the criteria suitable for this model [125]. Preliminary experimentation using different values for $k$ 10,15,20,25,30, with different pre-processing settings to cluster activity-related words, showed that twenty clusters was a good choice. The available data was investigated and clusters were manually examined; the value of $k$ was chosen where the daily activities set out in 3.3 emerged as natural groupings. As an attempt to explore the data, an initial experiment applied the elbow method on the available data. Although the effect was subtle, the resulting graph was consistent with the choice of $k$=20.

Euclidean distance is used here to measure the similarity between the vector representation of the words as shown in Formula 3.1:

$$d(w_j, c) = \sqrt{\sum_{i=1}^{200} (v_i^j - v_i^c)^2} \tag{3.1}$$

where $[v_1^j, ..., v_{200}^j]$ is the vector representing a new word to be added to a cluster, and $v^c$ is the vector representing the centroid of cluster c.

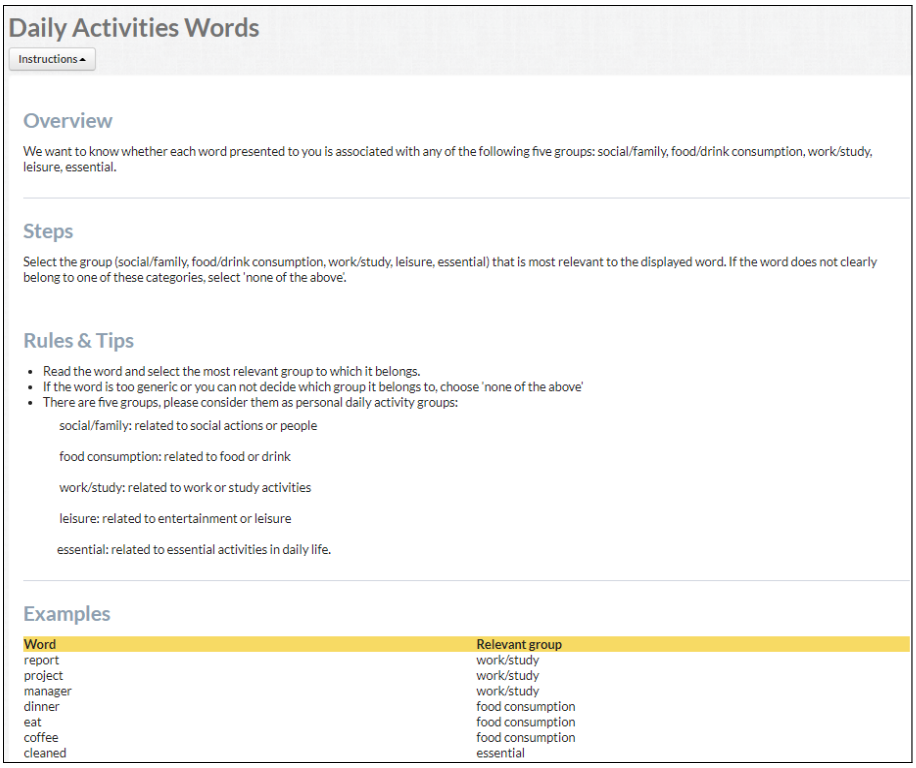
### 3.7.2 Word annotation

This step aims to give the clusters weights. Let $D$ be the set of words used in the diary data set (after pre-processing) with $D = \{w_1, ..., w_n\}$ and where each word $w_i$ was assigned a class $L(w_i)$ by the annotators where $L(w_i) \in L = \{work/study, social/family, food/drink, leisure, essentials, none\}$. For this annotation task, an online platform, Figure Eight (CrowdFlower)[2], was used where human annotators labelled the data.

[2] https://www.figure-eight.com

Each word was assigned to one activity group from the list: “work/study; social/-family; food/drink; leisure; essentials or none”. ‘None’ represents words that are not related to any of the activity groups. Each word was labelled by three different annotators following previous work [174, 175]. The annotation process underwent five steps:

1. **Data upload and management.** The data was uploaded in a tabular form and was prepared for the annotators to label.
2. **Job design.** The job title and overview were provided in addition to steps, tips, rules and examples to help with doing the jobs (Figure 3.6).



Daily Activities Words

Overview

We want to know whether each word presented to you is associated with any of the following five groups: social/family, food/drink consumption, work/study, leisure, essential.

Steps

Select the group (social/family, food/drink consumption, work/study, leisure, essential) that is most relevant to the displayed word. If the word does not clearly belong to one of these categories, select 'none of the above'.

Rules & Tips

- Read the word and select the most relevant group to which it belongs.
- If the word is too generic or you can not decide which group it belongs to, choose 'none of the above'
- There are five groups, please consider them as personal daily activity groups:

social/family: related to social actions or people

food consumption: related to food or drink

work/study: related to work or study activities

leisure: related to entertainment or leisure

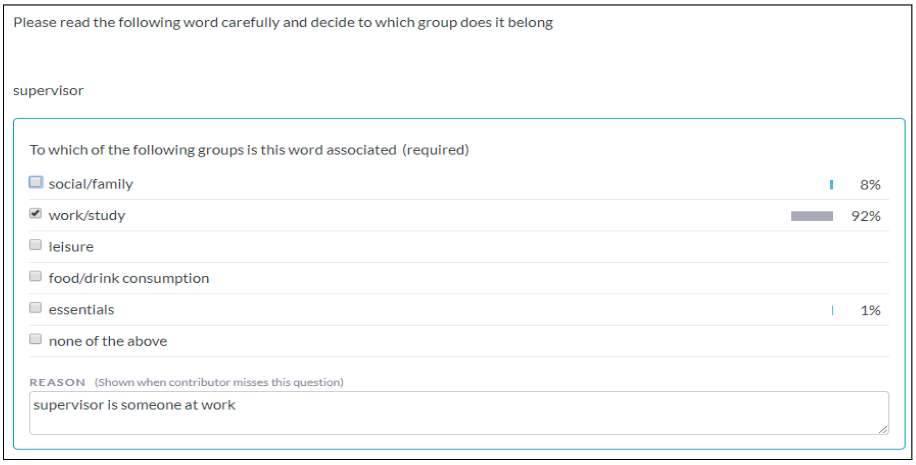
essential: related to essential activities in daily life.

Examples

| Word | Relevant group |
|---|---|
| report | work/study |
| project | work/study |
| manager | work/study |
| dinner | food consumption |
| eat | food consumption |
| coffee | food consumption |
| cleaned | essential |

Figure 3.6: Word annotation job design

3. **Quality testing.** This is an important quality control mechanism. Ideal test questions were created to ensure high-quality results from the job. Some data samples were provided with their labels for the contributors. The answers follow the rules that were given in the instructions and reasoning explains exactly how the answer was reached. Such questions are useful to test the contributors, to resolve any confusion and clarify why this is the right answer (Figure 3.7).



Please read the following word carefully and decide to which group does it belong

supervisor

To which of the following groups is this word associated (required)

□ social/family 8%

☑ work/study 92%

□ leisure

□ food/drink consumption

□ essentials 1%

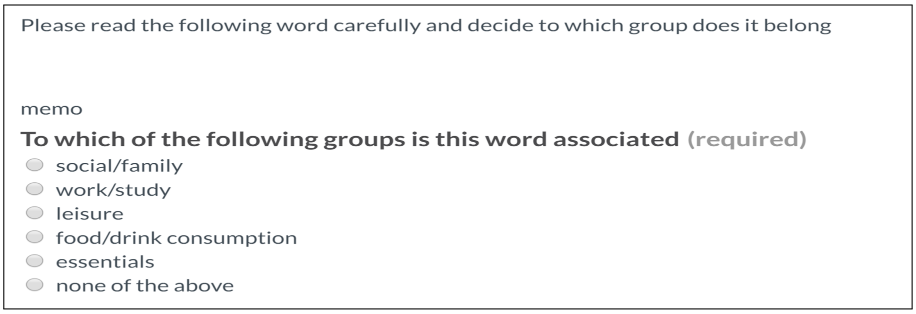
□ none of the above

REASON (Shown when contributor misses this question)

supervisor is someone at work

FIGURE 3.7: An example of word annotation quality testing

4. **Job launch.** The annotators were asked to read each word and choose the relevant activity group label (Figure 3.8).



Please read the following word carefully and decide to which group does it belong

memo

**To which of the following groups is this word associated** (required)

○ social/family

○ work/study

○ leisure

○ food/drink consumption

○ essentials

○ none of the above

FIGURE 3.8: Activity word annotation example

5. **Results.** This is where individual responses by the contributors as well as the aggregated results were provided in addition to other relevant information, such as the IP address, and contributors' ID. There were 250 contributors with an average confidence score of 0.9. 60% of the annotated entries were in complete agreement, 35.3% had two agreements and 11.7% had no agreement. Table 3.2 below reports a summary of the annotated data.

| **Activity** | work/study | social/family | food/drink | leisure | essentials | none |
|---|---|---|---|---|---|---|
| **% words** | 22.4 | 10.5 | 9.3 | 10.1 | 8.4 | 39.3 |

TABLE 3.2: Word annotations results from crowd-sourcing

### 3.7.3 The assignment of cluster weights

Each word has a label given by the annotators and belongs to a cluster based on its feature vector. Thus, it can be represented as (word, label, cluster). Accordingly, each cluster is assigned a weight based on the number/percentage of words related to the different activity groups in this cluster. Consequently, each word belonging to a cluster takes the weight of this cluster and can be represented as (word, cluster weight). The problem can be formulated as follows:
Let $C_1, ..., C_{20}$ be the partition of $D$ created in Section 3.7.1. For $l \in L = \{work/study, social/family, food/drink, leisure, essentials, none\}$, define a weight for cluster $C_i (1 \leqslant i \leqslant 20)$ as in Formula 3.2:

$$u_l^i = \frac{|\{w \in C_i : L(w) = l\}|}{|C_i|} \tag{3.2}$$

so that $u_l^i \in [0, 1]$ and $\sum_{l \in L} u_l^i = 1$

To generate the clusters' weights, three different approaches are presented here: including words labelled as 'none'; excluding words labelled as 'none' and replacing 'none' as 'essentials'. Each approach is explained below along with the results of the clustering weights.

1. **Including words labelled as 'none'**

   In this approach, all the annotated words were treated equally including 'none' where it represents a class in its own right. After pre-processing, there were still some words that may not belong to any of the groups (e.g., eventually, road, fire, area, came, etc.). Also, some entries may refer to a thought or represent a statement more than an activity. Although such entries are few, they still exist and should be taken into consideration here. Table 3.3 displays the weight of each cluster which was defined above in equation 3.2 as $u$. The classification approach build on these clusters are referred to as **"full activity classification"**, and results are explained in Section 4.4

| **Cluster** | **food %** | **essentials %** | **social %** | **leisure %** | **work %** | **none %** |
|---|---|---|---|---|---|---|
| cluster 1 | **90.16** | 1.64 | 1.64 | 0 | 1.64 | 4.92 |
| cluster 2 | 0 | 5.68 | 4.55 | 28.41 | 28.41 | **32.95** |
| cluster 3 | 0 | 3.49 | 5.81 | 3.49 | 13.95 | **73.26** |
| cluster 4 | 0 | 0 | 0 | 0 | 0 | **100.00** |
| cluster 5 | 0 | 0 | **91.49** | 0 | 2.13 | 6.38 |
| cluster 6 | 1.67 | 3.33 | 3.33 | 5.83 | **43.33** | 42.50 |
| cluster 7 | **41.18** | 23.53 | 0 | 5.88 | 3.92 | 25.49 |
| cluster 8 | 0 | 0 | 0 | 0 | 0 | **100.00** |
| cluster 9 | 1.72 | 3.45 | 4.31 | 6.90 | **50** | 33.62 |
| cluster 10 | 0 | 3.39 | 5.08 | 1.69 | **81.36** | 8.47 |
| cluster 11 | 0 | 1.14 | 17.05 | 20.45 | 14.77 | **46.59** |
| cluster 12 | 0 | 12.72 | 0 | 10.90 | 29.09 | **47.27** |
| cluster 13 | 4.65 | 11.63 | 6.98 | 5.81 | 25.58 | **45.35** |
| cluster 14 | 1.72 | 6.90 | 11.49 | 1.72 | 11.49 | **66.67** |
| cluster 15 | 23.33 | 8.33 | 8.33 | **26.66** | 11.66 | 21.66 |
| cluster 16 | 0 | 0 | 0 | 35.29 | 23.52 | **41.17** |
| cluster 17 | 0 | **64.29** | 0 | 14.29 | 0 | 21.43 |
| cluster 18 | 6.90 | 17.24 | 3.45 | 0 | 3.45 | **68.97** |
| cluster 19 | 2.33 | 27.91 | 6.98 | 13.95 | 2.33 | **46.51** |
| cluster 20 | 1.67 | 5 | 10 | 13.33 | 6.67 | **63.33** |

Table 3.3: Clusters and classes for words including ‘none’. The bold values denote the highest weights

2. **Excluding words labelled as ‘none’**

   In this approach, all words labelled with ‘none’ were removed and weights were given to the clusters to normalise. In the previous definitions, $C_i$ and $L$ are replaced by i.e. $C_i' = \{w \in C_i : L(w) \neq `none'\}$ and $L' = L - \{`none'\}$. Where it is assumed that these words may be irrelevant to the daily activities group and removing them increases the purity of the clusters to refer only to activities groups and later may increase classification accuracy. Table 3.4 displays the clusters weights. It should be noted in this approach that cluster 4 and cluster 8 were removed as they were pure ‘none’ clusters. The classification approach build on these clusters are referred to as **“non-trivial activity classification”**, and results are explained in Section 4.4

| Cluster | food % | essentials % | social % | leisure % | work % |
|---|---|---|---|---|---|
| cluster 1 | **94.83** | 1.72 | 1.72 | 0 | 1.72 |
| cluster 2 | 0 | 8.47 | 6.78 | **42.37** | **42.37** |
| cluster 3 | 0 | 13.04 | 21.74 | 13.04 | **52.17** |
| cluster 5 | 0 | 0 | **97.73** | 0 | 2.27 |
| cluster 6 | 2.90 | 5.80 | 5.80 | 10.14 | **75.36** |
| cluster 7 | **55.26** | 31.58 | 0 | 7.89 | 5.26 |
| cluster 9 | 2.60 | 5.19 | 6.49 | 10.39 | **75.32** |
| cluster 10 | 0 | 3.70 | 5.56 | 1.85 | **88.89** |
| cluster 11 | 0 | 2.13 | 31.91 | **38.30** | 27.66 |
| cluster 12 | 0 | 24.14 | 0 | 20.69 | **55.17** |
| cluster 13 | 8.51 | 21.28 | 12.77 | 10.64 | **46.81** |
| cluster 14 | 5.17 | 20.69 | **34.48** | 5.17 | **34.48** |
| cluster 15 | 29.79 | 10.64 | 10.64 | **34.04** | 14.90 |
| cluster 16 | 0 | 0 | 0 | **60** | 40.00 |
| cluster 17 | 0 | **81.82** | 0 | 18.18 | 0 |
| cluster 18 | 22.22 | **55.56** | 11.11 | 0 | 11.11 |
| cluster 19 | 4.34 | **52.17** | 13.04 | 26.09 | 4.35 |
| cluster 20 | 4.55 | 13.64 | 27.27 | **36.36** | 18.18 |

Table 3.4: Clusters and classes for words excluding ‘none’. The bold values denote the highest weights

What is apparent in this approach is that the activity group weights within the clusters have increased at different rates: 40% of clusters were dominated by work/study; 25% by leisure; 15% by essentials and only 10% by food/drink or social/ family activities.

3. **Replace all words labelled with ‘none’ as ‘essentials’**

In this approach all words labelled with ‘none’ were replaced as ‘essentials’ i.e. set $L(w) = `essentials'$ if $L(w) = `none'$. As the participants were required to record their daily activities, the assumption here is that all words should belong to daily activities, and to take advantage of all the words in the data set, the ‘essentials’ group definition is expanded to include any activity that does not belong to the other groups. Table 3.5 displays the clusters weights in this approach. The classification approach build on these clusters are referred to as **“essential classification”**, and results are explained in Section 4.4

In this approach, ‘essentials’ have clearly overcome the other activity groups in most of the clusters. Inspection of the other activities shows that there is at least one cluster that has one of the groups (food, social, and work) with the highest weight, except for ‘leisure’ which is never assigned with the highest value in any of the activities.

| **Cluster** | **food %** | **essentials %** | **social %** | **leisure %** | **work %** |
|---|---|---|---|---|---|
| cluster 1 | **90.16** | 6.56 | 1.64 | 0 | 1.64 |
| cluster 2 | 0 | 38.64 | 4.55 | 28.41 | 28.41 |
| cluster 3 | 0 | 76.74 | 5.81 | 3.49 | 13.95 |
| cluster 4 | 0 | 100.00 | 0 | 0 | 0 |
| cluster 5 | 0 | 6.38 | **91.49** | 0 | 2.13 |
| cluster 6 | 1.67 | 45.83 | 3.33 | 5.83 | 43.33 |
| cluster 7 | 41.18 | 49.02 | 0 | 5.88 | 3.92 |
| cluster 8 | 0 | 100.00 | 0 | 0 | 0 |
| cluster 9 | 1.72 | 37.07 | 4.31 | 6.90 | 50.00 |
| cluster 10 | 0 | 11.86 | 5.08 | 1.69 | **81.36** |
| cluster 11 | 0 | 47.73 | 17.045 | 20.45 | 14.77 |
| cluster 12 | 0 | 60 | 0 | 10.91 | 29.09 |
| cluster 13 | 4.65 | 56.98 | 6.98 | 5.81 | 25.58 |
| cluster 14 | 1.72 | 73.56 | 11.49 | 1.72 | 11.49 |
| cluster 15 | 23.33 | 30 | 8.33 | 26.67 | 11.67 |
| cluster 16 | 0 | 41.18 | 0 | 35.29 | 23.53 |
| cluster 17 | 0 | 85.71 | 0 | 14.29 | 0 |
| cluster 18 | 6.90 | 86.21 | 3.45 | 0 | 3.45 |
| cluster 19 | 2.33 | 74.42 | 6.98 | 13.95 | 2.33 |
| cluster 20 | 1.67 | 68.33 | 10 | 13.33 | 6.67 |

Table 3.5: Clusters and classes for words where ‘essentials’ replaces ‘none’

Overall, clustering diary words based on the features that were initially generated from the pre-trained model has proved to be useful in grouping personal activity words. The k-means clustering method, which has been commonly used in text mining research, returned acceptable results when applied to tweet word vectors, and generated activity clusters with acceptably high purity. Purity is an external evaluation criterion for cluster quality, it is a measure of the extent to which a cluster contains a single class. For example, *cluster 1* in Table 3.5, shows that ‘food’ is the majority class and the percentage of the members of this class in this cluster is 90%, which represent the purity of this cluster. Since the objective of this research was to investigate a broader framework rather than to find the best clustering method, a standard clustering approach is used as a step in the activity classification framework without seeking further improvements.

## 3.8 The classification of diary entries

Text classification is the process of assigning predefined categories to natural language text [176]. Here, each diary entry is automatically classified and given weights for the relevant activity groups. To implement the classification model, Python code was generated to check each entry word by word and assign it to the relevant class. The proposed algorithm starts by consulting the clusters by assigning each word to its relevant cluster and returning its cluster weight. It then adds up the weights for all the words in an entry and assigns the new weight to this entry as Figure 3.9 shows.

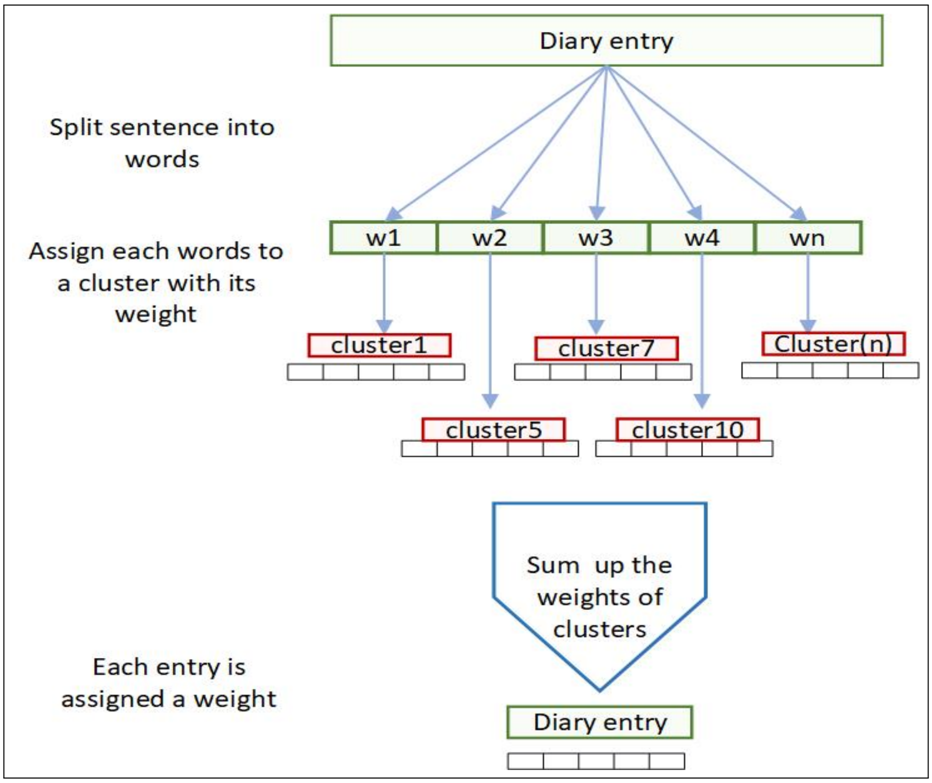

Diary entry
Split sentence into words
Assign each words to a cluster with its weight
w1
w2
w3
w4
wn
cluster1
cluster7
Cluster(n)
cluster5
cluster10
Sum up the weights of clusters
Each entry is assigned a weight
Diary entry

FIGURE 3.9: The diary entry activity classification framework

Given a diary entry $e = \{w_1, .., w_n\}$ where $w_i \in D\ \forall i$, let $C(w_i)$ where $(1 \leqslant i \leqslant n)$, be the cluster containing $w_i$, and let $W_a(C(w_i))$ be the weight of this cluster for activity $a \in L = \{work, social, food, leisure, essential, none\}$. Then the weight of entry $e$ for activity a, is given by Formula 3.3:

$$W_a(e) = \sum_{w \in e} W_a(C(w)) \tag{3.3}$$

Entry weight vectors were then normalised so that each $0 \leqslant W_a(e) \leqslant 1$ for $a \in L = \{work, social, food, leisure, essential, none\}$ and $\sum_{a \in L} W_a(e) = 1$
The activity of an entry is then set to be the activity with the highest weight as in Equation 3.4:

$$A(e) = \operatorname*{argmax}_{a \in L} W_a(e) \tag{3.4}$$

Figure 3.10 shows an example of an entry weight assignment referring to the cluster weights from Table 3.4 where 'none' are excluded. Such an instance would be classified as 'work' as the first class in cases of a single label classification and 'social' as the second class in cases of multi-label classification (see Section 4.6).

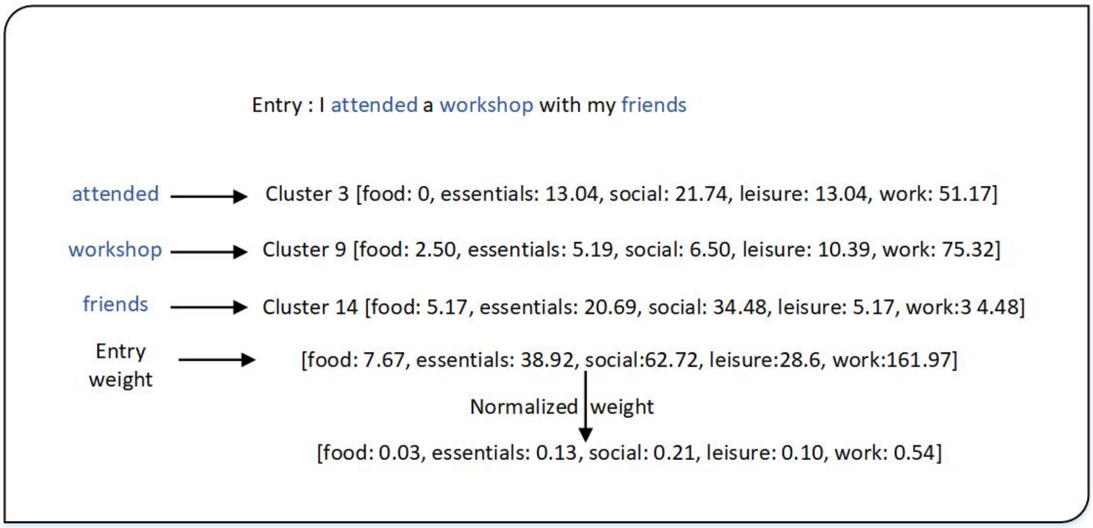

Entry : I attended a workshop with my friends
attended
Cluster 3 [food: 0, essentials: 13.04, social: 21.74, leisure: 13.04, work: 51.17]
workshop
Cluster 9 [food: 2.50, essentials: 5.19, social: 6.50, leisure: 10.39, work: 75.32]
friends
Cluster 14 [food: 5.17, essentials: 20.69, social: 34.48, leisure: 5.17, work:3 4.48]
Entry weight
[food: 7.67, essentials: 38.92, social:62.72, leisure:28.6, work:161.97]
Normalized weight
[food: 0.03, essentials: 0.13, social: 0.21, leisure: 0.10, work: 0.54]

FIGURE 3.10: An example for a diary entry classification

Table 3.6 displays some examples of entries and assigned weights. This sample is calculated consulting Table 3.3 where 'none' are included. As shown, some of the entries are highly relevant to some activity groups while others, such as the third entry, may be relevant to more than one group.

| **Entry** | **Weight** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **food** | **essentials** | **social** | **leisure** | **work** | **none** | $W_a(e)$ |
| Visiting husband's family | 0.04 | 0.02 | 0.65 | 0.08 | 0.06 | 0.16 | social |
| I met my supervisor at his office | 0 | 0.03 | 0.05 | 0.02 | 0.65 | 0.25 | work |
| I took my lunch in the office today | 0.30 | 0.02 | 0.04 | 0.02 | 0.32 | 0.29 | work |
| I drank my coffee in Nero | 0.74 | 0.09 | 0.01 | 0.02 | 0.02 | 0.12 | food |
| Off to bed | 0 | 0.64 | 0 | 0.14 | 0 | 0.21 | essentials |
| Chatting with my wife's grand mom | 0.02 | 0.04 | 0.63 | 0.02 | 0.10 | 0.20 | social |
| I had facial treatment and filling my lips but unfortunately, it is hurting me | 0.05 | 0.13 | 0.07 | 0.02 | 0.10 | 0.64 | none |

TABLE 3.6: Examples for entries activity-weights

In summary, the proposed framework for activity classification has overcome the weakness of the word embedding (lack of interpretability) and utilises the proximity vectors

(200 dimensions) to create contextual meaning through clustering. It generated lower-dimensional meaningful word probability vectors (6 dimensions) that clearly describe the activity a word refers to (explained in section 3.7), and thereby preserving proximity *effectively*.

## 3.9 Multi-label activity classification

When looking at diary instances, the situation often arises that more than one activity can be viewed as occurring simultaneously. For example, an entry such as "having lunch with my friend" could be classified as both 'food' and 'social' activities. This led to a multi-label problem which was defined by [177], where classes are not mutually exclusive and features may overlap.

As by definition, diary data can take more than one label at a time, technically each instance can be classified in any of the classes, but with different weights ranging from 0 to 1. Here, the labels with the two highest weights were assigned to each entry. Nevertheless, not all entries refer to more than one activity, some can only belong to one activity group such as the entry "having lunch".

For an entry $e$ whose activity been classified as $A(e)$ (as defined in Section 3.8), we define the second activity as $A^s(e) = \text{argmax}_{a \in L \setminus A(e)} W_a(e)$. Further analysis and evaluation are provided in Section 4.6.

## 3.10 Summary

This chapter has discussed the proposed framework for diary entries' categorisation into five daily activity groups: "work/study; social/family; food/drink; leisure and essentials". A word embedding based clustering-classification method was introduced here to solve the problem caused by a lack of training diary data for building a classifier by taking advantage of a large publicly available data source from Twitter as a starting point. The framework is composed of four modules. First, all entries were transformed into words and went through customised *pre-processing* steps to filter out the non-meaningful words to the domain. A *pre-trained word embedding model* was then utilised to generate word-features vectors. Based on these vectors, the data was grouped using k-means *clustering* algorithm, and weights were then given to the clusters based on annotation. Lastly, for *classification* each entry was broken down to its contents and checked for its belonging to a predefined cluster. Then, entries are assigned a vector of activity-weights and classified to the relevant activity group accordingly. This framework generates short feature vectors for all entries (using Equation 3.3) that allows a novel approach to diary text classification based on tweets word embedding clustering. The next chapter provides a detailed evaluation of this framework and comparisons to other states of the art methods.

# Chapter 4

# Classification analysis

## 4.1 Introduction

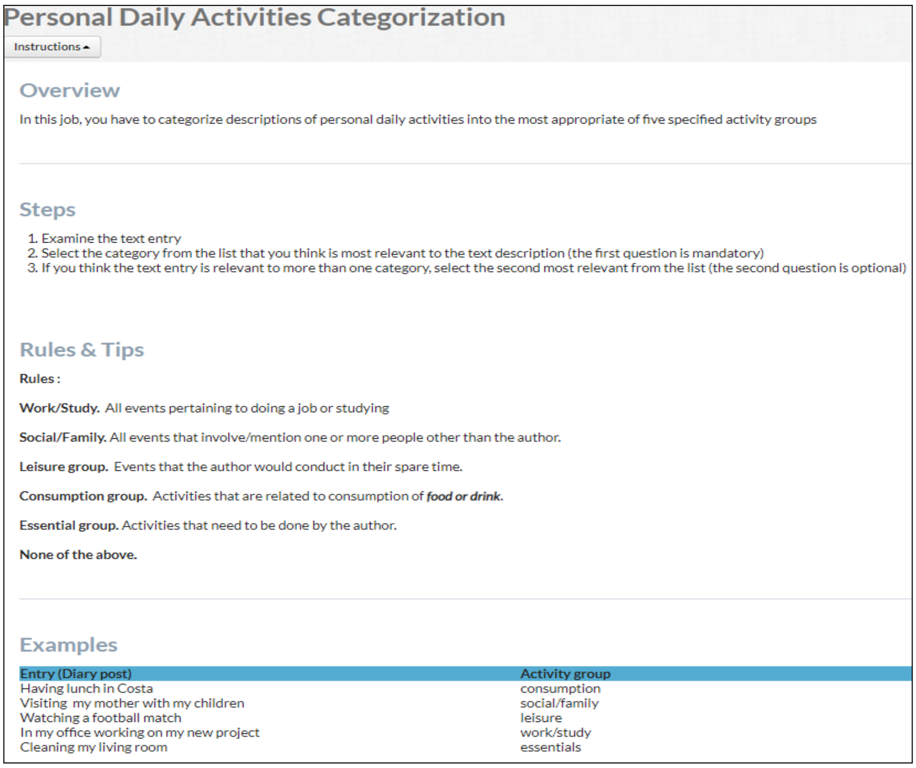
The previous chapter described the proposed framework for the classification of diary activities, and showed how a textual diary entry is given a weight that represents its relevant activity group. This chapter provides an evaluation of this framework for classifying daily activities based on the collected data described in Section 3.2, it also reports on the results for the three different approaches presented in the previous chapter and presents different views and levels of analysis.

Details of the steps taken are provided in the following six subsections: Section 4.2 establishes a ground truth through a manual annotation process to validate the results. Sections 4.3 presents the metrics used to evaluate the classifier's performance. Section 4.4 provides an evaluation of the activity classification within the three different approaches. To handle 'none', Section 4.5 investigates the effect of using different thresholds and evaluates the classification task overall and for each activity group. In Section 4.6, the classifier performance in detecting entries with multi-labels is evaluated. Closing with a comparison between the proposed classifier and three alternative text classifications algorithms in Section 4.7.

## 4.2 Ground truth

Manual categorisation was used here to evaluate the performance of the proposed classification model on the diary data set, this was done through the online crowd sourcing platform Figure Eight (previously CrowdFlower). This service provides high quality results and has been previously used for short text categorisation and evaluation for similar applications [178–180]. In this approach human annotators were asked to label the data. The process has five steps, which are similar to those in Section 3.7.2 but differ in the job design, aims, objectives and instructions.

1. **Data upload and management.** The data was uploaded in tabular form and prepared for the annotators to read and label.
2. **Job design.** In this stage the following were provided: the job title and overview; the steps required to do the job; tips and rules, as shown in Figure 4.1.



Personal Daily Activities Categorization

Overview

In this job, you have to categorize descriptions of personal daily activities into the most appropriate of five specified activity groups

Steps

1. Examine the text entry
2. Select the category from the list that you think is most relevant to the text description (the first question is mandatory)
3. If you think the text entry is relevant to more than one category, select the second most relevant from the list (the second question is optional)

Rules & Tips

**Rules :**

**Work/Study.** All events pertaining to doing a job or studying

**Social/Family.** All events that involve/mention one or more people other than the author.

**Leisure group.** Events that the author would conduct in their spare time.

**Consumption group.** Activities that are related to consumption of ***food or drink***.

**Essential group.** Activities that need to be done by the author.

**None of the above.**

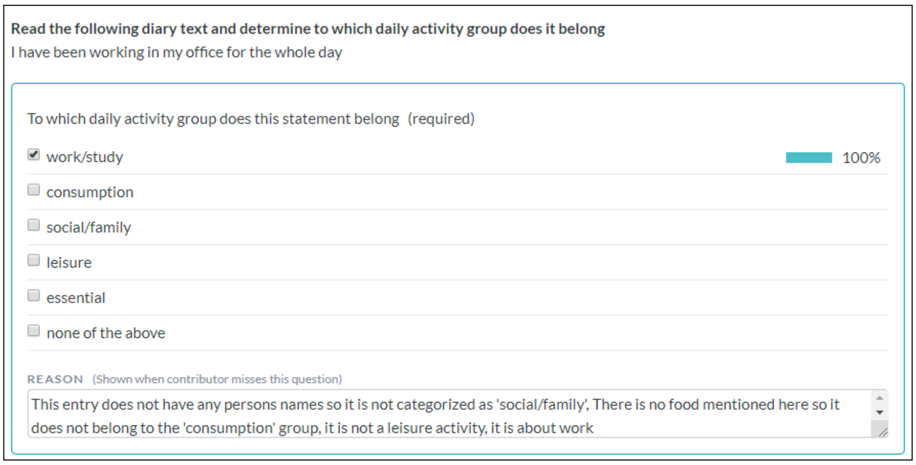
Examples

| Entry (Diary post) | Activity group |
|---|---|
| Having lunch in Costa | consumption |
| Visiting  my mother with my children | social/family |
| Watching a football match | leisure |
| In my office working on my new project | work/study |
| Cleaning my living room | essentials |

FIGURE 4.1: Activity classification job design

3. **Quality testing.** This is an important quality control mechanism where ideal test questions were created to ensure high quality results from the job. Specifically,

these are samples of the data with their ideal labels for the contributors. The answer followed the rules that were given in the instructions and reasoning which explained exactly how the answer was reached. Such questions are useful to test the contributors, to resolve any confusion and to clarify why this is the right answer. An example is shown in Figure 4.2.



**Read the following diary text and determine to which daily activity group does it belong**
I have been working in my office for the whole day

To which daily activity group does this statement belong (required)

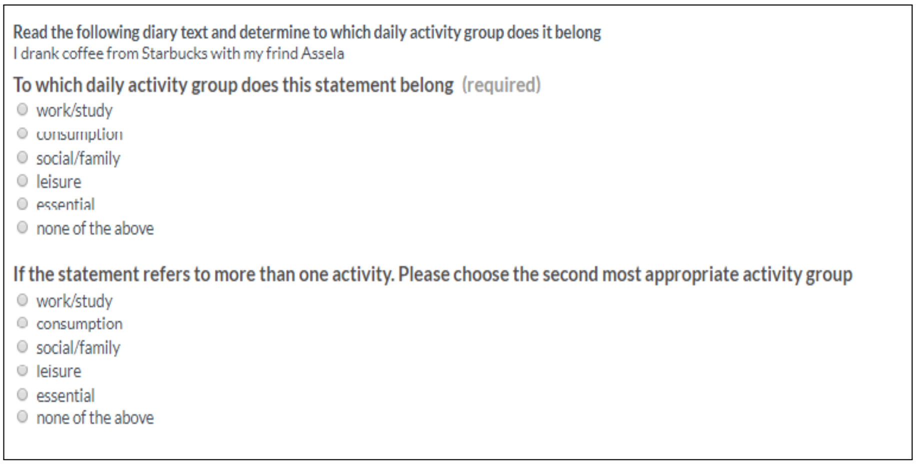
☑ work/study 100%
□ consumption
□ social/family
□ leisure
□ essential
□ none of the above

REASON (Shown when contributor misses this question)

This entry does not have any persons names so it is not categorized as 'social/family', There is no food mentioned here so it does not belong to the 'consumption' group, it is not a leisure activity, it is about work

Figure 4.2: Activity classification quality questions

4. **Launching the job.** The annotators were asked to read each diary entry and label the data with the relevant daily activity category from a given list of choices. An example is shown in Figure 4.3.



**Read the following diary text and determine to which daily activity group does it belong**
I drank coffee from Starbucks with my frind Assela

**To which daily activity group does this statement belong** (required)

○ work/study
○ consumption
○ social/family
○ leisure
○ essential
○ none of the above

**If the statement refers to more than one activity. Please choose the second most appropriate activity group**

○ work/study
○ consumption
○ social/family
○ leisure
○ essential
○ none of the above

Figure 4.3: A diary entry classification example

5. **Results**. This is where individual responses by the contributors were given, along with other relevant information like IP address, contributors ID and trust score. In total, there were 401 contributors with an average confidence score of 0.97.

### 4.2.1 Gold standard

The annotated data represents the ground truth that will be used to evaluate the automatically classified data. Humans are known to differ in their opinions and judgements, for example an entry like “having my morning coffee” could be interpreted by some people as ‘food/drink’, whereas by others it could be interpreted as ‘leisure’. So, each entry was labeled by three annotators as recommended by the technical support team of CrowdFlower and following previous work [181], three human judgments should be sufficient to get the average agreement for the best result and reach sound judgment for each entry. The agreement rates for the entries were as follows: for the first label those that had full agreement accounted for 62%; entries that had two annotators agreeing accounted for 31% and those with no agreement accounted for 7%. For the second label: entries with full agreement accounted for 56%; 38% for two annotators' agreement and 6% for no agreement. It should be noted that the first question was required, each entry was assigned with at least one label, about 60% of the entries had exactly one label.

In the diary data set $D = \{e_1, ..., e_n\}$, each entry $e_i$ was assigned a primary class $L(e_i)$ by the annotators and potentially a second class $L^s(e_i)$, where $L(e_i), L^s(e_i) \in \{food, work, social, leisure, essentials, none\}$. Each $e_i$ was assigned with a confidence score that was calculated as the average of the three agreements that were obtained from the annotators. Each entry was assigned with the label with the highest confidence score as the first label. Only entries that had two or three agreements were counted, and those with only one agreement were considered as ‘none’. Table 4.1 below shows the distribution of the ground truth annotation over the six categories.

Table 4.2 displays some examples of the annotated data

| Annotation | First label total | First label % | Second label total | Second label % |
|---|---|---|---|---|
| Food/drink consumption | 476 | 19.6% | 39 | 3.9% |
| Work/study | 616 | 25.4% | 122 | 12.3% |
| Social/Family | 451 | 18.6% | 358 | 36.0% |
| Essentials | 147 | 6.1% | 59 | 5.9% |
| Leisure | 265 | 10.9% | 348 | 35.0% |
| None | 468 | 19.3% | 72 | 7.2% |
| total | 2423 | 100% | 995 | 100% |

Table 4.1: Distribution of annotations in the gold standard

| Example | First label | Second label |
|---|---|---|
| I attended a workshop with my friends | work/study | social/family |
| I cleaned my flat | essentials | none |
| I had my coffee with my friend at Starbucks | food/drink | social/family |
| I watched a movie at home | leisure | none |

Table 4.2: Examples of annotations in the gold standard

## 4.3 Evaluation measures

In text classification, the performance of a classifier is evaluated by testing its ability to correctly identify the relevant class of a given entry. Given a testing data set, the effectiveness of the classifier is measured relevant to the training set by producing four values, which in the context of this work are:

**True Positive (TP)** is the number of entries that are correctly classified, i.e., the label of the ground truth activity class matches that assigned by the classifier, e.g., “visiting my uncle”, its actual activity group is ‘social’ and it is classified accordingly.

**True Negative (TN)** is the number of entries that are correctly classified as negative (unrelated activity entries), e.g., “having lunch” does not belong to the ‘social’ activity group; therefore, it is not classified in this group.

**False Positive (FP)** is the number of entries that are incorrectly classified, e.g., “I called my sister this morning” is classified as ‘work’, although its actual group is ‘social’.

**False Negative (FN)** is the number of entries that are falsely classified as unrelated activity entries, e.g., any activity that is related to, for example, ‘social’ but classified in any of the other groups.

These four counts of the number of entries that are correctly and incorrectly classified for each activity group constitute a confusion matrix, which is shown in Table 4.3

| | | Predicted activity group | |
|---|---|---|---|
| | | Related activity group | Non-related activity group |
| Actual activity group | Related activity group | TP | FN |
| | Non-related activity group | FP | TN |

Table 4.3: A confusion matrix

The performance of the activity classification model is evaluated by three common metric measures in text classification: precision; recall and F-measure.

**Precision** (Formula 4.1) expresses the number of cases that the model says were relevant and that were relevant. It is the ratio of correctly classified positive observation to the total classified positive entries

$$Precision = \frac{TP}{TP + FP} \tag{4.1}$$

**Recall** (Formula 4.2) is the ability of a model to correctly identify all the data cases in a class. It is the ratio of correctly classified positive observations to all observations in the actual class, i.e., the fraction of all positive patterns that are correctly classified.

$$Recall = \frac{TP}{TP + FN} \tag{4.2}$$

**F-measure** (Formula 4.3) is the weighted average of precision and recall. It is the harmonic mean of precision and recall taking both metrics into account. It gives equal weight to both measures and represents the optimal balance between the two measures.

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.3}$$

Usually, the goal is to maximise the value of these measures. In activity classification, precision, recall and F-measure have different levels of importance in different circumstances, i.e., choosing the right measure depends on the application. For example, if the aim was to predict people with mental health, sleeping or eating disorders based on their daily activities, then recall value may be the most important measure; however, if the purpose of the application is to obtain restaurant suggestions, then precision should be prioritised.

## 4.4 Activity classification: a comparison of approaches

In the diary data set $D = \{e_1, ..., e_n\}$, each entry $e_i$ was assigned a primary class $A(e_i)$ by the classification framework, where $A(e_i) \in L = \{food, work, social, leisure, essentials, none\}$, and $A(e_i)$ is the activity class with the maximum weight/probability $W_a(e_i)$ as explained in Section 3.8.

Section 3.7.3 proposed three classification approaches. The classification results for these approaches are reported in Tables 4.4, 4.5, and 4.6. The classifier performance was evaluated in terms of precision, recall and F-measure. The Python Scikit-learn (sklearn.metrics) machine learning library was used to calculate these measures.

The first approach, which we term **"full activity classification"**, included 'none' as a distinct class. This method considers the possibility that some diary entries may refer to a thought or a statement rather than an activity, this came from having words that do not belong to any of the activity groups (results are shown in Table 4.4).

| **Activity group** | **Precision** | **Recall** | **F-measure** |
|---|---|---|---|
| Food/drink | 89% | 96% | 92% |
| Work/study | 94% | 92% | 93% |
| Social/Family | 98% | 62% | 76% |
| Essentials | 42% | 17% | 24% |
| Leisure | 87% | 11% | 20% |
| None | 80% | 99% | 88% |

Table 4.4: Precision, recall, F-measure for full activity classification

The second approach is where 'none' is excluded as a possible class, and is referred to as **"non-trivial activity classification"**. This method assumes that all diary entries are activity-related as participants were instructed (Table 4.5).

The third approach where all the 'none' words were replaced with essentials, referred to as **"essentials activity classification"**. The aim here was to take advantage of all the words and consider the assumption that all entries belong to activities, so the essential

| Activity group | Precision | Recall | F-measure |
|---|---|---|---|
| Food/drink | 84% | 96% | 90% |
| Work/study | 54% | 98% | 69% |
| Social/Family | 94% | 75% | 83% |
| Essentials | 29% | 54% | 38% |
| Leisure | 64% | 70% | 67% |

Table 4.5: Precision, recall, F-measure for non-trivial activity classification

group definition was expanded here to include any activity that does not belong to the other activity groups (Table 4.6).

| Activity group | Precision | Recall | F-measure |
|---|---|---|---|
| Food/drink | 94% | 95% | 94% |
| Work/study | 95% | 88% | 91% |
| Social/Family | 98% | 42% | 59% |
| Essentials | 11% | 100% | 20% |
| Leisure | 0% | 0% | 0% |

Table 4.6: Precision, recall, and F-measure for essential activity classification

There is a clear difference in classifying the activity groups between the approaches. To examine this further, two methods are explored here: analysing the clusters and the individual activity classes.

### 4.4.1 Cluster analysis

Clustering was a foundational step carried out to assign word weights with respect to activities. As entry classification was built on these clusters, this section investigates how word clusters affected classification. The inspection of the clusters showed that some of them are clearly defined clusters, as they are strongly associated with one class. For example, as can be seen from Tables 3.3, 3.4, and 3.5, 'food/drink' only appeared in half of the clusters of this data set and usually with very high weights (sometimes 90%-95%). Also, 'social/family' reached 92%-98% and 'work/study' reached 81%-89%, that may sometimes cause them to overcome any of the added weights of the other words in an entry and accordingly are classified more correctly.

In other cases, there are no clusters that are most heavily associated with any class. This can be seen clearly in Table 3.5, where there is no leisure-dominant cluster. The reason for this is that ‘essentials’ exceeded the other activity groups in most of the clusters, except for ‘leisure’ which was never assigned with the highest value in any of the clusters. Therefore, it is not very surprising that the classifier failed to detect any of the entries in this group, as shown in Table 4.6. Another observation is that when ‘none’ were excluded, as shown in Table 3.4, 40% of the clusters became strongly associated with the ‘work/study’ class, which has led to the increase in recall for this class, as shown in Table 4.5.

### 4.4.2 Individual activity classification analysis

In all three approaches, some activities seem to be easier to detect than others. In particular, entries that belong to the classes ‘food/drink’, ‘work/study’ and ‘social/-family’. This is probably because the definitions of these groups are clear; especially ‘food/drink’, where the classifier performance is high in all three approaches.

Looking more closely at the ‘work/study’ group, it is not surprising that the recall score is the highest of all other activity groups when ‘none’ was removed (see Table 4.5), this can be attributed to the fact that words related to this group dominated other classes, as explained earlier. However, for the same reason, precision was relatively low: 46% of the returned data related to other groups. This resulted in the lowest F-measure weight of this class when compared with the other approaches.

When ‘none’ was replaced by ‘essentials’, as noticed from Table 4.6, the ‘social/family’ F-measure dropped significantly when compared with the other two approaches. Although the precision of this group reached 98%, the recall was very low: almost 60% of the relevant entries were not detected. A possible explanation for this is that ‘essentials’ dominated most of the clusters and probably a number of words in these clusters also appeared in the ‘social/ family’.

Unlike the above classes, the definition of the ‘essentials’ class is not as precise; hence, in terms of F-measure, the classifier performance when detecting entries in this group is low in the three approaches. When ‘none’ was included, precision was higher than recall, and inspection revealed most of the returned entries were relevant to the activity ‘sleep’, such as : “Off to bed”; “I will have a nap now”; “trying to wake up” and “going to sleep”. When ‘none’ was excluded, recall became higher than precision. Interestingly, when inspecting the returned entries this activity group embraced more varied examples, such as:" I washed my clothes"; "I unpacked my luggage" and "got up and took a shower". One anticipated finding was when 'none' was replaced as 'essentials', all instances relevant to this activity group were detected; however, the ability of this model to identify only the relevant instances became weak, with 89% of instances wrongly classified to this group. This clearly caused confusion in the classification task, that being due to the fact that although many new relevant entries were detected, such as "I cleaned my place" and "finally shaving at the barber shop", many irrelevant instances were classified as belonging to this group such as " I went to the cinema with my friend", as this category represented the overwhelming majority of the records.

The classifier performance varies between the three approaches when detecting entries belonging to the 'leisure' group, the best results were when 'none' was removed, the recall score was 70%, and precision was relatively high: 64%. While when 'none' was included the precision was high, only 13% of the returned entries were irrelevant to this activity group. Furthermore, the ability of the classifier to identify the relevant entries was very low: 11%. It is reasonable to assume that this happened because of the normalisation when removing 'none' words which increased the 'leisure' weights. And as mentioned earlier, the classifier could not correctly classify any of the entries in this group when 'none' was replaced with 'essentials' as the number of 'essentials' words increased and the weight of this group exceeded 'leisure' in all clusters, which resulted in labelling a high number of entries as 'essentials' instead of 'leisure'.

## 4.5 Activity classification with different thresholds

In an attempt to improve the performance of the classifier, i.e. achieve the best precision vs recall balance, the proposed method is evaluated with different thresholds $\tau$ from 0.1 to 0.9. The threshold is altered for classifying the positive cases is as set out below.

In the diary data set $D = \{e_1, ..., e_n\}$, each entry $e_i$ is assigned a class $A(e_i)$ where $A(e_i) \in L = \{food, work, social, leisure, essentials, none\}$, with the classification changed to 'none' if the weight is below $\tau$, defined in Formula 4.4:

$$A(e) = \begin{cases} A(e) & W_{A(e)}(e) \geq \tau \\ none & \text{otherwise} \end{cases} \tag{4.4}$$

Essentially, entries where there is insufficient support for any non-trivial class are classified as 'none'. Such instances are referred to as "implicit none"; whereas, the instances that were originally classified as 'none' based on their given weight are referred to as "explicit none". Accordingly, there are different perceptions of the classifier behaviour when identifying instances in the 'none' class in the three approaches. In the first case, 'none' presents a class in addition to any instances with weight under a threshold which became members of this class (explicit and implicit 'none'). The other two cases only contain implicit 'none' entries, with any instances with weight under a specific threshold being labelled as 'none'.

### 4.5.1 Full activity classification (including 'none' as a class)

Table 4.7 shows that as the threshold increases, recall and precision decrease. In general, (in a binary classifier or multi-class classifier with different classification method), it is expected, due to the tradeoff between precision and recall, that precision decreases when recall increases. So, these results appear counter-intuitive. However, this can be explained, as instances with a weight lower than the threshold are re-labelled with 'none'. As a result, the overall precision and recall decrease for all activity classes; except the 'none' class, which behaves differently to the other groups thereby affecting the performance measures, especially precision as Figure 4.4 shows.

| **Threshold** | **Precision** | **Recall** | **F-measure** |
|---|---|---|---|
| 0.1 | 85% | 85% | 82% |
| 0.2 | 85% | 85% | 82% |
| 0.3 | 85% | 83% | 81% |
| 0.4 | 76% | 77% | 73% |
| 0.5 | 68% | 65% | 60% |
| 0.6 | 62% | 50% | 41% |
| 0.7 | 62% | 39% | 27% |
| 0.8 | 61% | 34% | 21% |
| 0.9 | 46% | 32% | 19% |

Table 4.7: Precision, recall, F-measure for full activity classification with different thresholds

Three general observations can help to interpret these results as the threshold increases. First, true positives decreased in all groups. Second, except for 'none', false negatives increased. Third, again except for 'none', false positives decreased.

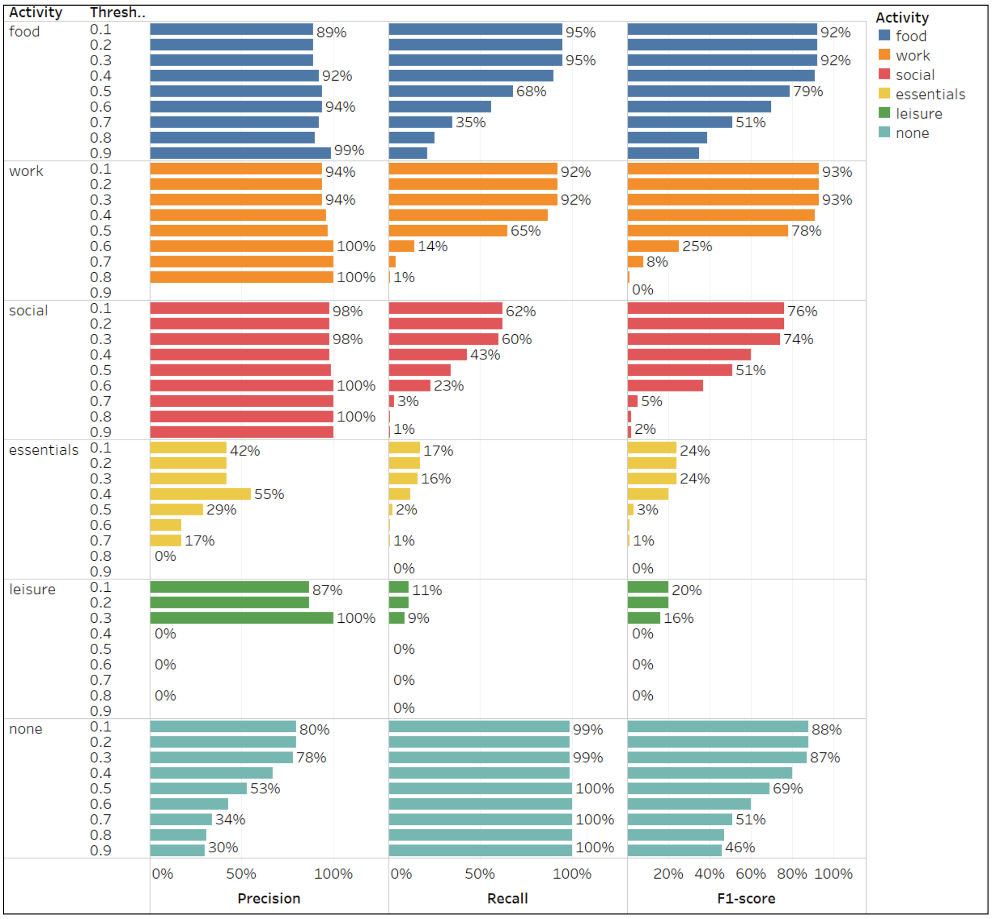


Figure 4.4: Precision, recall, F-measure for full activity classification with different thresholds

Recall decreased with increasing thresholds for all activity groups except for 'none', which is discussed later. 'Food/drink' and 'work/study' reached very high recall for low thresholds ($\tau$ <=0.4), decreasing with increasing thresholds to reach 0 with $\tau$=0.9 in the 'work/study' class. With the 'social/family' group, the classifier behaves in the same manner as the previous two classes; however, the highest recall score (62%) is relatively low. With the 'leisure' and 'essentials' groups, recall scores were very low, and entries were not detected with high thresholds, since there were no entries with high weights in these groups; nevertheless, it decreases slightly with increasing thresholds. Surprisingly, the precision rate slightly decreases with increased threshold, despite neither the number of true positives, nor the number false positives decreasing in all groups. The balance between these two rates caused the decrease in precision to be very slight, this is especially evident for 'food/drink', 'work/study' and 'social/family'.

The precision of the ‘leisure’ and ‘essentials’ groups classification also decreases, but because the few members of these groups have very low weights, the classifier was unable to detect any instances with $\tau >= 0.3$ for ‘leisure’. Nevertheless, all the returned instances belong to this group since no irrelevant instances were classified to be members of this group. Finally, ‘essentials’ displayed unusual behaviour, as it increased with an increasing threshold, but with $\tau = 0.4$ the true positive rate increases where more than half of the instances are detected (from 2% to 12%) and the false positive rate doubled (from 5% to 10%), so the precision score was the highest.

Regarding the ‘none’ class, it is important to note here that entries are those which were explicitly classified to be ‘none’ by the classifier and are in addition to entries implicitly labelled as ‘none’. Here, the classifier behaves oppositely to the other groups. When increasing the threshold, false negatives decreased and false positives increased; accordingly, recall increased and precision decreased. The recall is always high; therefore, the classifier can always find the relevant instances in this group, which were explicitly labelled as ‘none’ and this value slightly increases with increasing threshold as implicit ‘none’ appear. Precision, on the other side, decreases with the increase of the threshold; this is expected because the false positives increases, since the classifier re-labels instances from the other groups with ‘none’ with increasing thresholds. The performance here is not only affected by explicitly labelled ‘none’ but also implicitly labelled ‘none’, which caused precision to decrease. This means that some of the instances that belong to their activity groups are mistakenly classified to be ‘none’ when they are under a threshold.

### 4.5.2 Non-trivial activity classification (excluding ‘none’ as a class)

Table 4.8 reports the classification results in the case of excluding ‘none’. The rule here differs from the previous approach when using different thresholds. The performance was the lowest in terms of F-measure (25%) with the highest threshold 0.9. It improved gradually with decreasing threshold until it reached the best performance with threshold 0.4, then it dropped again with decreasing threshold. Also, with $\tau = 0.5$, the measurement scores are very close to those from 0.4 where recall is the same (72%), but precision is lower by only 4% points.

| Threshold | Precision | Recall | F-measure |
|---|---|---|---|
| 0.1 | 55% | 66% | 59% |
| 0.2 | 55% | 66% | 59% |
| 0.3 | 71% | 66% | 60% |
| 0.4 | 74% | 72% | 71% |
| 0.5 | 70% | 72% | 69% |
| 0.6 | 67% | 62% | 57% |
| 0.7 | 61% | 45% | 39% |
| 0.8 | 60% | 38% | 29% |
| 0.9 | 60% | 35% | 25% |

Table 4.8: Precision, recall, F-measure for non-trivial activity classification with different thresholds

As Figure 4.5 shows, recall always decreases with increasing threshold, compared with the result in the previous approach where 'none' were included, the classifier performance when detecting 'food/drink' and 'work/study' is the same, and it performs better when identifying 'social/family' activities, the highest score is 74%. Interestingly, the classifier performance is much better with the 'leisure' and 'essentials' groups: with 'leisure', recall score reaching 70% and 'essentials' 54%.

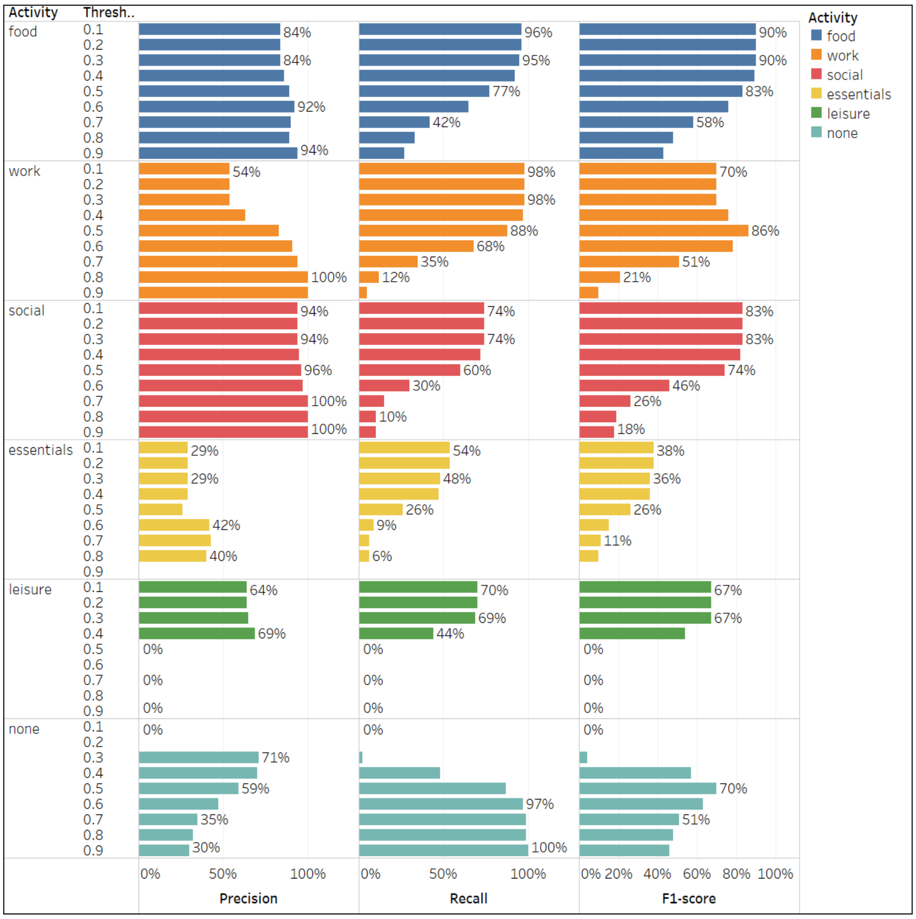


Figure 4.5: Precision, recall, F-measure for non-trivial activity classification with different thresholds

Precision increased with increasing threshold. For 'work/study' there is a notable decrease from $\tau < 0.4$ which explains and supports the findings from the cluster analysis that the number of clusters associated with this activity increased.

With 'leisure' and 'essentials' the precision is lower than the previous approach; however, the classifier began to detect the relevant instances accurately at higher thresholds. Regarding 'none', there were no instances initially classified as 'none', this is because with this approach words labelled with 'none' were removed. However, instances are implicitly classified to be members of this group and re-labelled with 'none' when they have weights under a specific threshold. The classifier starts to detect 'none' with $\tau = 0.3$ (the lowest weights were 0.25) and recall increased gradually with increasing threshold. This means when increasing the threshold, the number of true positive increased and false negatives decreased. So, the classifier can detect more instances related to this group even though they were not originally classified as ‘none’. Precision, on the other hand, decreased with increasing thresholds since false positives increased. This is to be expected, since as threshold increases more irrelevant instances may be classified as ‘none’.

### 4.5.3 Essentials activity classification (replacing ‘none’ as ‘essentials’)

Table 4.9 shows the classifier performance measures evaluated with different thresholds when ‘none’ words were replaced by ‘essentials’. The best classifier performance was when $\tau = 0.5$ where F-measure was 49%.

| Threshold | Precision | Recall | F-measure |
|---|---|---|---|
| 0.1 | 47% | 42% | 40% |
| 0.2 | 47% | 42% | 40% |
| 0.3 | 47% | 42% | 40% |
| 0.4 | 62% | 43% | 44% |
| 0.5 | 62% | 47% | 49% |
| 0.6 | 59% | 46% | 40% |
| 0.7 | 59% | 38% | 27% |
| 0.8 | 60% | 28% | 17% |
| 0.9 | 48% | 24% | 14% |

Table 4.9: Precision, recall, F-measure for essentials activity classification with different thresholds

Figure 4.6 shows the classification performance measures for each activity group. It is clear that classifier performance deteriorated when predicting some of the classes in this approach.

Looking at recall, the classifier’s ability to detect relevant instances in ‘food/drink’ and ‘work/study’ is similar to the previous approaches. With ‘social/family’ recall was weak; more than half of the relevant instances were missed, even with the lowest threshold. With ‘essentials’, although F-measure decreased, the recall was high in this approach it was expected, since the number of instances in this group increased and as the words in this group increased; accordingly, their cluster weights increased. This caused the classifier to fail to detect any of the instances in the ‘leisure’ group as explained earlier. Precision, for ‘food/drink’, ‘work/study’, and ‘social/family’ was similar to that of the

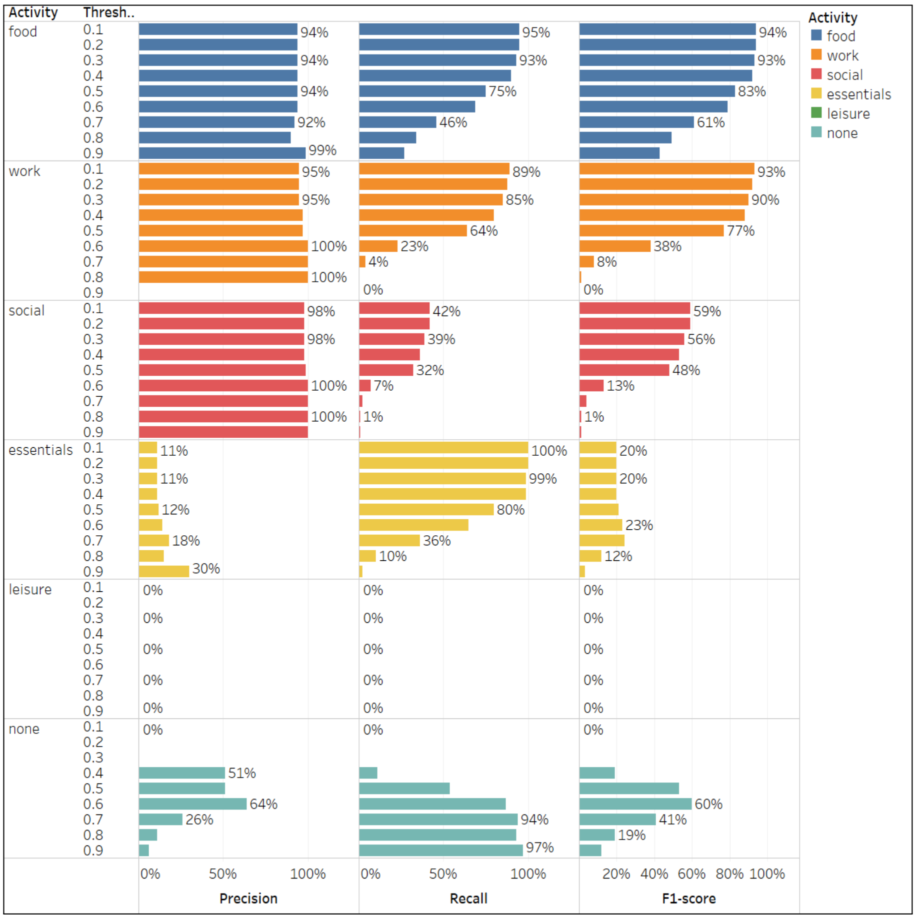


FIGURE 4.6: Precision, recall, F-measure for essentials activity classification with different thresholds

previous approaches, it was consistently high with a slight decrease with decreasing thresholds. There were no instances labelled with 'leisure' and the precision scores for detecting 'essentials' were low, the highest was 30% with $\tau = 0.9$.

When it comes to 'none', similar to the previous case, the classifier was able to detect instances in this group, even though they were not initially classified to be members of this group. The classifier started to detect implicit 'none' with $\tau = 0.4$ and increased until it reached 97% with $\tau = 0.9$. Precision decreases with decreasing thresholds, whilst the highest score was with $\tau = 0.6$.

## 4.6 Multi-label activity classification

In determining which of the entries have one label and which have two labels in the data set: *Firstly*, the annotated data is examined where each instance has at least one label and, at most, two labels. Then, the labels' cardinality is calculated, which is the average number of labels of instances in the data set [182, 183] as shown in Formula 4.5.

$$LCard = \frac{1}{N}\sum_{i=1}^{N} \mid L_i \mid \tag{4.5}$$

where $N$ is the number of instances in the diary data set $D$ and $L_i$ is the set of labels for $i$th instance assigned by the annotators and $L_i \subset L$ ={food/drink, social/family, work/study, essentials, leisure,none} the set of all possible activity labels. The calculated cardinality was 1.4 for $N$=2,423 where 1,428 instances have one label and 995 have two labels.

*Secondly*: a threshold was calibrated from the weights of the second labels of the classified data such that $LCard\{\text{classified data}\} \approx LCard\{\text{annotated data}\}$. So, thresholds were found for the three different approaches where the label cardinality of the classified data was close to 1.4. Accordingly, as explained in Section 3.9, each entry $e_i$ is assigned to one or two classes $A(e_i)$ and $A^s(e_i)$, where:

$$e_i \text{ has} \begin{cases} multilabel & \text{if } A^s(e_i) \geq \tau \\ singlelabel & \text{otherwise} \end{cases} \tag{4.6}$$

Table 4.10 shows the thresholds and number of instances for each classification approach.

| Approach | Threshold | Number of instances |
|---|---|---|
| Full activity classification | 0.27 | 1068 |
| Non-trivial activity classification | 0.25 | 970 |
| Essentials activity classification | 0.28 | 1250 |

Table 4.10: Accuracy for multi-label classification for different approaches

For the multi-label evaluation design, precision, recall, accuracy and F-measure were calculated for each approach using Equations 4.7, 4.8, 4.9 and 4.3 respectively:

$$Precision = \frac{1}{N}\sum_{i=1}^{N}\frac{\mid L_i \cap \{A(e_i) \cup A^s(e_i)\} \mid}{\mid A(e_i) \cup A^s(e_i) \mid} \tag{4.7}$$

$$Recall = \frac{1}{N}\sum_{i=1}^{N}\frac{\mid L_i \cap \{A(e_i) \cup A^s(e_i)\} \mid}{\mid L_i \mid} \tag{4.8}$$

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{\mid L_i \cap \{A(e_i) \cup A^s(e_i)\} \mid}{\mid L_i \cup \{A(e_i) \cup A^s(e_i)\} \mid} \tag{4.9}$$

where, $A_i$ and $A_i^s \subseteq L$ are the set of labels generated by the classifier.

Table 4.11 displays the results for the multi-label evaluation for three proposed approaches. Precision, recall, F-measure and accuracy were calculated using the above equations.

| **Approach** | **Metric** | | | |
|---|---|---|---|---|
| | Precision | Recall | F-measure | Accuracy |
| Full activity classification | 77.6% | 26.0% | 39.0% | 71.3% |
| Non-trivial activity classification | 53.8% | 10.8% | 18.0% | 42.3% |
| Essential activity classification | 47.1% | 18.8% | 26.9% | 34.3% |

Table 4.11: Results for multi-label activity classification

Table 4.12 shows the results for the single-label design for the three proposed approaches. It displays the overall weighted averages of precision, recall, F-measure and accuracy from the three earlier result Tables 4.4, 4.5 and 4.6.

| **Approach** | **Metric** | | | |
|---|---|---|---|---|
| | Precision | Recall | F-measure | Accuracy |
| Full activity classification | 85.2% | 84.7% | 82.3% | 84.7% |
| Non-triveal activity classification | 55.0% | 66.0% | 58.5% | 66.0% |
| Essential activity classification | 46.6% | 41.7% | 40.0% | 41.7% |

Table 4.12: Results for single-label activity classification

From Tables 4.11 and 4.12, it can be concluded that in both, the single-label and the multi-label designs, the classifier performance was the best with the “full activity classification” approach when ‘none’ was included and represented a distinct class. Comparing the two methods shows that single-label classification returned higher results in all three approaches than the multi-label. Precision is higher than recall with both designs. However, there was a significant difference between the two designs in terms of recall (and accordingly, F-measure). The classifier’s ability to detect the relevant activities was better with the single-label design, while it returned low recall scores with all three approaches with the multi-label design, indicating the poor performance of the classifier when detecting classes of multi-label entries. Also, accuracy, which is the percentage of correctly classified entries, was better with the single-label classification method. In both methods, the highest score was with the “full activity classification” approach and the lowest score with the “essential activity classification” approach. However, F-measure (precision and recall) provides more insight into a classifier’s functionality than accuracy and is not sensitive to changes in the data distribution, especially where there are multiple classes.

One of the factors that may have affected the multi-label classification design results is the different number of entries that represent the data set in each approach as a result of different thresholds, unlike with single-label where the number of entries is the same in the three approaches.

This has led to two design choices. Firstly, the “full activity classification” is the selected approach to be used in the rest of this research, as the performance was the best in both cases. This corresponds to the gold standard, where 19.3% of the entries were annotated to be ‘none’ as shown in Table 4.1. Secondly, the most suitable and precise way to represent the activities in the diary entries is by the activity probability vectors generated from Equation 3.3, as they show the exact probability of each activity in an entry.

## 4.7 Activities classifier evaluation: a comparison with other classifiers

To evaluate the performance of the proposed model as a whole working framework, it was tested against three machine learning algorithms that have been proven to work well for text categorisation [184], these are: Support Vector Machine (SVM); Decision Tree (DT), and Naive Bayes (NB).

**SVM** represents the state of the art of many classification tasks, including events extraction from personal generated data [120] and mining short texts [121]. It is a supervised discriminative algorithm that is characterised by producing a hyper-plane which separates the data into different classes and derives the widest channel between the classes [122]. It can be used for linear classification, as well as non-linear classification using the kernel function [185]. The main advantage of this algorithm is that it is very resilient to over-fitting.

**DT** is a classifier in the form of a structured tree, as its name implies. Each node can be either a leaf or a decision node. A leaf shows the class label, while the decision node implies a test to classify the data. Internal nodes are labelled by attributes and branches departing from them are labelled based on the weight that an attribute has in the text document. It categorises the data by recursively testing the weights of the attributes and labelling internal nodes until a leaf is reached. It can easily be tracked from the input (root)to the output (leaf) and needs less training time since it does not need any parameter settings [122].

**NB** was chosen here because it performs well with multiple classes in textual data. It is a probabilistic classifier that measures the probability of a document to be a member of a category and deals with each one independently. It has proven to be a powerful approach for many reasons: it is simple, easy to implement, fast to compute and efficient when compared with non-Naive Bayes approaches, this is because of the independency [124].

Recurrent neural network (RNN) architectures also have been shown to return strong results in text processing. However, work in a similar context to categorise diary text into activities by Bremer et al. [65], showed that the usage of the RNN approach (compared to BoW) does not improve but rather decreases the classification performance. The main reason suggested was that the training data (diary text) were not sufficient or accurate enough for the RNN to generate the knowledge and connection between the words and categories. Therefore, this approach might have only added noise in such classification.

Weka [186, 187] was used to implement the classification experiment. It consists of a collection of machine learning algorithms for data mining. All text diary entries were converted into feature vector using a BOW representation. A ten-fold cross-validation was implemented on the gold standard created in Section 4.2.1 for the training of and testing of the different classifiers. The labelled data set was split into ten equal groups and trained ten times. Each time, nine groups were used for training and the remaining one was used for testing. Table 4.13 shows the results of the comparison between the proposed model and the three classification models with the highest performance: SVM, DT, and NB. The overall performance represents weighted average results.

| **Classifier** | **Precision** | **Recall** | **F-measure** |
|---|---|---|---|
| Proposed Model | 85.0% | 85.0% | 82.0% |
| Support Vector Machine | 83.2% | 52.2% | 52.2% |
| Decision Tree | 82.4% | 46.5% | 43.9% |
| Naive Bayes | 82.3% | 45.0% | 41.8% |

Table 4.13: Precision, recall, F-measure, and accuracy for three different classifiers and the proposed model

As shown in the table, the proposed model returned the highest F-measure score of 82%. Precision was high with all models and the differences between scores being slight. However, it is clear that the ability of detecting the relevant instances was best with the proposed model (recall=85%), when compared with the other models: the proposed semi-supervised model outperformed the other three supervised models, especially in terms of recall.

The proposed model is compared against these three algorithms as examples of the types of methods that could be applied to a BoW approach in a simple and a direct way. However, note that finding the best BoW approach is not the aim of this research and each of these could potentially be improved by further refinement. There is a range of different approaches that could be tested to classify the data, these representative approaches were reasonable choices as they showed that the proposed approach is better across the board than the others.

## 4.8 Summary of findings and conclusion

This chapter provided a detailed analysis and evaluation of the framework for classifying diary entries into five activity groups: food/drink, work/study, social/family, leisure, and essentials. Entries were represented as low dimensional probability feature vectors that indicate the relevant activity. Each entry was given a weight based on its words. Each word gained its weight by combining its vector space (its places in a different space: tweets) and label given by human annotators, that were originally given to the cluster of which it belongs. Overall, performance was good, although there are interesting differences between the different activity classes, evaluation metrics and design choices in the approaches.

This chapter proposed and compared three different approaches to handling trivial or uninteresting diary entries in the activity classification task. The first view was to consider 'none' as a class in its own right, testing the ability of the classification task to differentiate between the activity groups, assuming that a diary includes both activity-related and non-activity-related entries (e.g., thoughts). The second view was to exclude 'none' as a class in an attempt to encourage the classification algorithm to assign entries to the closest activity group, with an implicit assumption that diary entries are all activity related. The third view was to assume all words labeled with 'none' belong to the 'essentials' group, using a broader definition that includes any routine activities that do not fit in any of the other groups.

Including 'none' as a class in its own right was the best approach since it returned better classification results, not only for 'none' but also the more clearly defined groups ('food/ drink', 'work/ study', 'social/family'). The only groups with a lower performance level were 'essentials' and 'leisure'.

As expected, excluding 'none' as a distinct class increased recall for all activity groups. However, this also increased confusion between classes, with precision notably decreased for all activity groups. Although, this approach could be used when recall is the required measure, it was dismissed for the remainder of this thesis since there will always be entries that do not refer to activities.

Finally, replacing 'none' with 'essentials' was not successful. Although the results were close to those of the first approach (where 'none' was included); increasing the capacity of this group to take in more diverse activity included too many irrelevant entries.

Moreover, it limited the performance regarding finding entries in the 'leisure' group.

The decision whether to include, exclude, or replace 'none' was important, as analyses of the data from these three possible views showed that each has its own advantages and disadvantages. As including 'none' is more rational and has returned acceptable results, the results from this method are used in the remainder of this work.

This study also found that the nature of classes plays an important role in classification. This was apparent with 'food/drink', 'work/study' and 'social/family', which were clearly defined. 'Food/drink' is taken as an example as it had high prediction results in the three approaches. In this data set it can be seen that the percentage of words was around 9% (see Table 3.2) in this group and did not affect the way they were clustered. They were aggregated based on their intrinsic features (200-dimension word embedding vectors), which matched their extrinsic features (words annotation). This explains the presentation of this group in clusters, as it only appears in half of the clusters and usually with a very high percentage. This is in contrast to 'leisure' which had a close percentage to 'food/drink' (around 10%), the highest weight in the clusters were between 28-42% which are comparatively low. However, this means that the intrinsic and extrinsic features do not complement each other, as this group is not well-defined.

Setting the same thresholds for all groups seems to be unfair. The weights of the instances in each group varies because of the clusters and is highly dependent on the data set itself. In the first approach where 'none' represents a distinct class, the best classifier performance came with the lowest thresholds. Thus, using different thresholds did not make any difference to the results; the rule is that the lower the threshold, the better the performance. However, this is not the case with the other two approaches, where in the case of excluding 'none', the best classifier performance was when $\tau$ =0.4 (Table 4.5) and when 'none' was replaced with 'essentials' the classifier performed the best with $\tau$ =0.5 (Table 4.9). Therefore, it is favourable to have different thresholds not only for each approach, but also for each activity group within each approach; since each activity group has different weights and they behave differently with different thresholds.

Finally, when considering multi-labelled data, the classifier performance was the best when 'none' represented a class in its own right and confirmed that accuracy is also the best with this approach.

Comparing the proposed classification method returned better results than some common classification algorithms, such as SVM, DT and NB, especially in terms of recall and improved daily activities classification by taking advantages of multiple semi-supervised learning methods.

# Chapter 5

# Predicting emotion

## 5.1 Introduction

All individuals experience changes to their emotions on a daily basis, these changes can play a crucial role in their wellbeing. There are many ways one can express emotions; they can be communicated verbally, by facial expression, gestures, actions, tones, and written text. As discussed earlier in this thesis, a diary is a common medium for the recording of a person's daily activities, as much as a diary is rich with informative data describing daily events, it is also rich with attitudinal information about emotional states.

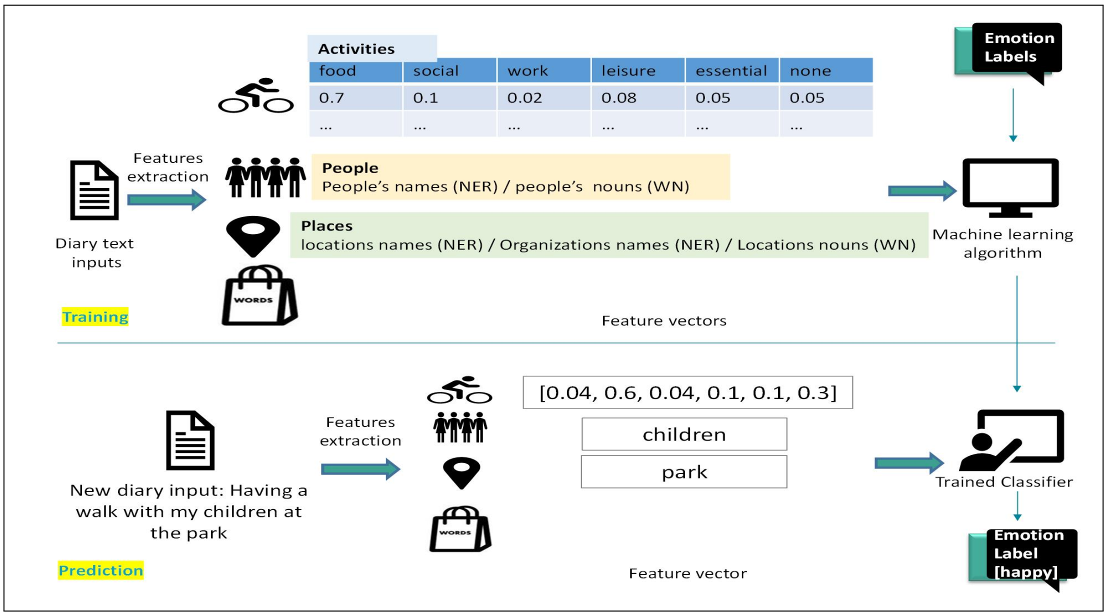
This chapter investigates the feasibility of inferring emotions from a diary using domain-specific features, namely daily activities in association with people and places as representations of the individuals in the context of real life. As short text does not provide sufficient word occurrence and traditional classification methods such as Bag-of-Words have limitations. In an attempt to address this problem, a range of differentiating features exhibited by diarists, which had relatively high accuracy with the activities classification are exploited to enhance emotion prediction.

Emotion is an essential indicator of mental health. For example, research in psychology relates social anxiety with major depressive disorder [188]. Also, emotion is a crucial factor in behavioural studies, where for example, emotion disorder may be associated with an absence from school [189]. Automating the process of feature extraction and emotion prediction is therefore very useful as it will give more insight from the data, reduce the time and effort for analysis, and provide a good recommendation system to undertake/avoid activities to manage emotions like anger or excitement.

Predominantly, this chapter looks at different aspects of the prediction of emotions, including comparing different machine learning strategies for training the classifiers: personal and global, using different performance evaluation measures, and comparing emotions prediction performance in the context of three different models (Ekman's basic emotion, the Circumplex along with a simpler pleasantness/unpleasantness classification). The two emotions schemes, Ekman's and Circumplex (which were discussed in Section 2.3.4), were selected based on a study [190] which investigated the performance of supervised machine learning in emotion prediction and provided an experimental justification for the choice of emotions classification schemes in free text. They were found to be the best schemes with the highest F-measure values in a comparison between six different emotions schemes. A simpler third emotion scheme of pleasantness/unpleasantness emotions is proposed here; the motivation of which was the previous classification investigation and was derived from the Circumplex model. For each scheme, emotions distribution is discussed, the results as a whole are investigated, followed by individuals' dis-aggregated data analyses and, finally, the application of statistical significance tests to validate the results.

## 5.2 Features

In order to train a classifier, each diary entry is represented by a vector of features. The proposed predictive model takes these feature vectors that represent the context as input and produces an output of an emotional state (e.g. happy) an individual may be in that context. Figure 5.1 shows the prediction process for a diary entry.



| Activities | | | | | |
|---|---|---|---|---|---|
| food | social | work | leisure | essential | none |
| 0.7 | 0.1 | 0.02 | 0.08 | 0.05 | 0.05 |
| ... | ... | ... | ... | ... | ... |

FIGURE 5.1: Emotion prediction process

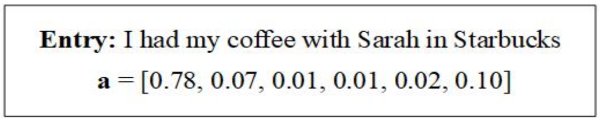
Here, classic features that are used in various types of classification (BOW) in addition to diary domain-specific features (activities, people, and places) are used to train the classifier. These features are explained in detail below.

### Bag of Words (BOW)

All textual diary entries are represented by their words as features where the order is not considered, but only the presence of the words with the exclusion of stop words. Vectors, whose features are derived from the occurrence of these words are generated.

### Activities weight-vectors

These are numerical vectors that were generated earlier in Section 3.8. For an entry $e_i$, there is a vector of label weights = $[l^i_{food}, l^i_{social}, l^i_{work}, l^i_{leisure}, l^i_{essentials}, l^i_{none}]$ is built by adding the words weights of that entry. Each $0 \leqslant l^i_a \leqslant 1$ for $a \in A = \{food, social, work, leisure, essential, none\}$ and $\sum_{a \in A} l^i_a = 1$.



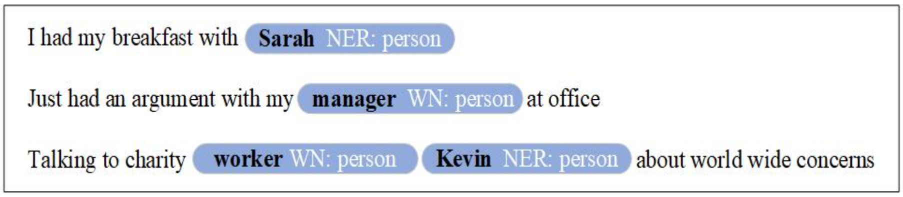
**Entry:** I had my coffee with Sarah in Starbucks

**a** = [0.78, 0.07, 0.01, 0.01, 0.02, 0.10]

FIGURE 5.2: An example for an activity vector

### People

People mentioned in one's diary are likely to have participated or be relevant to the described event. Thus, people names (e.g., Sarah, Mathew, etc.) and roles (e.g., teacher, doctor, etc.) are extracted from the text and added to the features vector for emotions prediction. To automatically identify people in diary texts, NER and WordNet are used here. NER is used to extract people names, while the WordNet lexicographical file (noun.person) is consulted where the algorithm checks for the words that represent human roles.



I had my breakfast with **Sarah** NER: person

Just had an argument with my **manager** WN: person at office

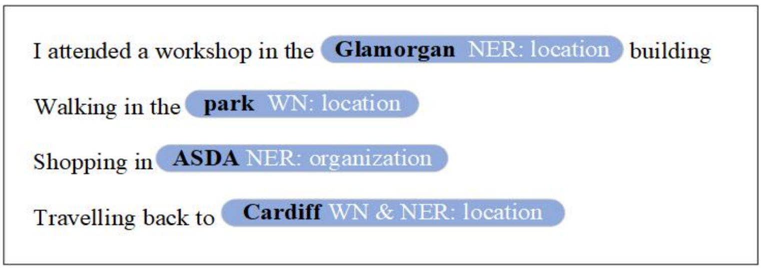
Talking to charity **worker** WN: person **Kevin** NER: person about world wide concerns

Figure 5.3: Examples for people extraction

People names and roles can be important features for classification. Mentioned roles are likely to be associated with specific emotions and most probably present in a relevant activity. For example, an entry including a mention of a family member may be labelled with an emotion like 'happy' in a 'social/family' activity. Similarly, when people mention roles associated with a work related activity, such as a supervisor or clients, they probably will be associated with an emotion. Such features are common and they share similar associations (meaning) amongst all people. Therefore, they are used with both global and personal training (as will be explained later). Whilst, people names have different associations between individuals. Thus, in a single person's diary there is a set of names that may be associated with some emotions and other factors like places and activities and would not be helpful to use such features in global training. Therefore, they are only considered in personal training.

### Places

Places that individuals visit also affect their emotions in different ways. Here, places are locations that people mention in their text which vary from very granular like home, park, and gym to high level like a city or a country. To automatically identify locations expressions in diary texts, NER and Wordnet are used. WordNet Lexicographical file (noun.location) is consulted to detect locations such as 'home' and 'park'. NER is used to detect locations such as 'Costa' or 'Waterloo'. Also, organisations such as 'Starbucks' and 'Tesco' are added as people mention them as places. It should be noted here that organisations and locations detected by NER are complementary to each other. Also, both NER and WordNet detect countries or cities' names, for example Cardiff, UK, Egypt, etc.



I attended a workshop in the **Glamorgan** NER: location building

Walking in the **park** WN: location

Shopping in **ASDA** NER: organization

Travelling back to **Cardiff** WN & NER: location

FIGURE 5.4: Examples for places extraction

The same applies to people. Since locations extracted using WordNet are common between people and their presence with similar kinds of activities may be associated with emotion; accordingly, this could help with prediction; whether with global or personal training. While with NER, a set of location entities that describe places are associated with each person's diary are used as features to predict emotions in personalised training only.

## 5.3 Experimental design

This section explains the data preparation, the classifier choice and the different learning strategies for emotion prediction.

### 5.3.1 Data preparation

As described in Section 3.2, participants were given the option to choose an emotion word from one of Ekman's emotions list, or enter their own. Analysing the data revealed that there are around 110 different emotion words which participants used to label their entries, since they were able to enter their emotions as free text. Table 5.1 displays the words and their frequencies in this data set. Interestingly, although users were given a drop-down list of Ekman's emotions, in 30% of the cases, users entered their own emotions (70% of entries that were neither happy nor neutral provided their own emotion). So, for the purpose of classification, these words were mapped to two different emotion classification schemes Ekman's and the Circumplex using the crowdsourcing approach as done in previous work [191, 192]. By running an emotion-words mapping job in CrowdFlower, where each word was annotated by three annotators who had to choose from separate lists of the two emotion schemes to map the word. They were selected to be located in countries that the first language is English with a high level of confidence and the average trust score was 0.95. All words annotated with two or three agreement were considered. Only 5% had no agreement, and they were labelled as 'neutral'. The aggregated results of all participants are presented in Sections 5.5 and 5.12.

| **Emotion** | **Count** | **Emotion** | **Count** | **Emotion** | **Count** | **Emotion** | **Count** |
|---|---|---|---|---|---|---|---|
| **neutral** | 745 | inspired | 4 | terrified | 1 | crazy ! | 1 |
| **happy** | 692 | full | 4 | relieved | 1 | emotional | 1 |
| excited | 194 | confused | 4 | irritated | 1 | extremely happy | 1 |
| **sad** | 92 | stressed | 4 | sympathetic | 1 | so happy | 1 |
| upset | 67 | sick | 4 | dizzy | 1 | so upset | 1 |
| tired | 64 | enthusiastic | 4 | overwhelmed | 1 | giggling | 1 |
| content | 29 | intrigued | 4 | confident | 1 | depressed | 1 |
| relaxed | 25 | **surprised** | 4 | addled | 1 | hot | 1 |
| bored | 20 | strssed | 4 | deeply disappointed | 1 | pretty | 1 |
| good | 18 | enjoying | 4 | frustrated | 1 | enjoy | 1 |
| sleepy | 15 | nostalgic | 3 | ecstatic | 1 | moved | 1 |
| peaceful | 13 | Interested | 3 | cheery | 1 | mixed feelings | 1 |
| worried | 12 | fun | 2 | joyful | 1 | indulged | 1 |
| satisfied | 11 | calm | 2 | pleased | 1 | moved and tearful | 1 |
| **fearful** | 10 | entertained | 2 | blessed | 1 | reminiscing | 1 |
| **angry** | 10 | euphoric | 2 | mixed emotions | 1 | contemplating | 1 |
| mad | 10 | nervous | 2 | creative | 1 | cheerful | 1 |
| boredom | 10 | laughing | 2 | energetic | 1 | sentimental | 1 |
| grateful | 9 | headache | 2 | anxiety | 1 | sympathy | 1 |
| fine | 9 | amazed | 2 | soo excited | 1 | obligated | 1 |
| lazy | 9 | pain | 2 | in love | 1 | unhappy | 1 |
| anxious | 8 | thinking | 2 | glad | 1 | unimpressed | 1 |
| exhausted | 6 | focused | 1 | not bad | 1 | drained | 1 |
| hungry | 6 | impassive | 1 | sooo good | 1 | agitated | 1 |
| disappointed | 6 | unsettled | 1 | shocked | 1 | refreshed | 1 |
| hopeful | 5 | discouraged | 1 | scared | 1 | **disgusted** | 0 |

Table 5.1: Emotion word-frequencies. Ekman's emotions are the ones in bold

### 5.3.2 Classifier choice and training approaches

To select an algorithm for classification, different learning methods were initially tested on the data set, with SVM outperforming the other methods. This was also consistent with the investigation in [190] to classify emotions in texts from different schemes, which found that SVM returns the best results. SVM, therefore, is used in the chapter for emotion classification. However, for completeness, the discussion section in 5.8 of this chapter also provides a comparison with other prediction methods.

The models are trained with a set of data instances where emotion is known, with each instance, a feature vector and an emotion label, i.e., happy, sad, etc. Precision, recall and F-measure of the outputted predictions were calculated for evaluation. Two different machine learning strategies were followed for training and testing the classifiers.

**Personalised:**

In this approach, the model is trained and tested on a single individual's data. Taking different feature sets: activities vectors referred as **Act**, activities; people and places referred to as **APP**, **BOW**, a combination of activities and BOW referred as **Act-BOW,** and a combination of activities, people, places and BOW referred to as **APP-BOW**. The model is then tested using 10 fold cross-validation. Noting that 'people' in this personalised model include two features: people's proper names (e.g. Sarah) and people's common nouns (e.g. friend). Also, 'places' include proper places names (e.g. ASDA) and common places referees (e.g. home).

**Global:**

While personalised models return good results, they require individual training and they need a bigger data set generated by the same person for a long period of time. To reduce the amount of training, a global emotion model is proposed here. It is created by an aggregation of all of the participants' data. This model can be used as an initial model for a new person, bootstrapping the training procedure. The model is tested using unseen data where a user's data (used as a testing set) is removed from the all-users' data which is (used as the training set) following the procedure in [66, 193]. About 90-94% of the data was used for training and 6-10% for testing, these percentages varied according the each participant's data set size. As with the personalised model, the global model is trained by the same feature sets. However, 'people' and places here only refer to the common nouns (e.g. supervisor, park).

## 5.4 Ekman's emotion scheme

This model has six emotions: *happiness; sadness; anger; fear; surprise and disgust*, together with *neutral* to represent the absence of a clear emotional state.

### 5.4.1 Emotions distribution

The distribution of reported emotions across all participants for Ekman's model is displayed in Figure 5.5, with the aggregate shown as P0, which shows that participants were generally happy or in neutral states as they occupy the highest percentage of the labelled entries. In this data set, 'disgust' is an emotion that no one used to express their feeling in their diaries, and 'surprise' was used infrequently by a few participants. The figure also shows the emotions distribution in each participant's data, which highlights the wide variations between them, and that some emotions are not expressed by all, as with 'surprise' and 'fear'.

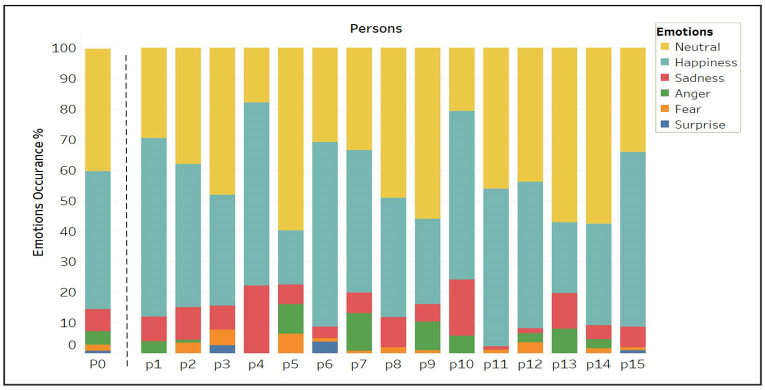


FIGURE 5.5: Ekman's emotions distribution

### 5.4.2 Prediction results analysis

As stated earlier, two learning strategies were followed: *personalised* and *global*. Table 5.2 displays the prediction results for the personalised model of the three approaches using activities, people and places as features (APP), Bag-Of-Words (BOW), and a combination of both approaches (APP-BOW). Table 5.3 displays the prediction results for the global model using activities (Act), Bag-Of-Words (BOW), and a combination of both approaches (Act-BOW).

The cells highlighted in pink indicate the higher values between personalised and global models. Showing that personalised training mostly performed better especially for the negative emotions classes.

It is apparent from the tables that 'happy' and 'neutral' are the two most well-predicted emotions. Specifically, predicting 'neutral' had the highest recall with Act and APP in

| | APP | | | BOW | | | APP-BOW | | |
|---|---|---|---|---|---|---|---|---|---|
| **Emotion** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** |
| neutral | 50.5 | **63.2** | 55.0 | **55.9** | 62.4 | **58.7** | 54.9 | 62.6 | 58.3 |
| happiness | 56.4 | 60.3 | 55.7 | **64.7** | 64.4 | **63.9** | 63 | **64.6** | 63.4 |
| sadness | 10.6 | 2.4 | 3.9 | 27.8 | 14.4 | 17.7 | **37.2** | **21.2** | **25.3** |
| fear | 0 | 0 | 0 | **18.2** | **13.3** | **15.4** | **18.2** | 11.0 | 13.6 |
| anger | 13 | 2.4 | 3.9 | **39.4** | **13.1** | **19.2** | 28.9 | 10.3 | 14.5 |
| surprise | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 5.2: Ekman's personalised prediction. Bold values indicate the highest values across approaches

| | Act | | | BOW | | | Act-BOW | | |
|---|---|---|---|---|---|---|---|---|---|
| **Emotion** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** |
| neutral | 52.3 | **67.8** | **57.1** | **54.3** | 52.9 | 51.5 | 53.5 | 55.2 | 52.4 |
| happiness | **58.8** | 61.3 | 57.5 | 56.4 | **72.3** | 61 | 58.0 | 70.3 | **61.5** |
| sadness | 0 | 0 | 0 | 9.2 | 10 | 8.2 | **10.1** | **11.5** | **9.7** |
| fear | 0 | 0 | 0 | 3.0 | 1.5 | 2.0 | 3.0 | 1.5 | 2.0 |
| anger | 0 | 0 | 0 | **14.3** | **5.4** | **6.9** | 13.9 | **5.4** | 6.7 |
| surprise | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 5.3: Ekman's global prediction. Bold values indicate the highest values across approaches

both personalised and global models. 'Happiness' has highest recall using APP-BOW with personalised and using BOW with global model. Apparently, *people use similar language in describing their activities in their 'happiness' state.*

In the personalised model, 'sadness', 'fear', and 'anger' were poorly predicted with higher precision compared with recall. This is still better than the global model, where the classifier could not predict any of the negative emotions using activities as features on their own. The reason could be that these classes consist of very few entries compared to the global training data set. Another reason might be that *there is variation in activities that cause unpleasant emotions while there may be some similarities in activities related to 'happiness' or non-emotional (neutral) entries.*

Entries labelled with 'surprise' could not be predicted with any of the three approaches. One possible reason is that they represent a very small percentage of the data (only found in three participants data). Another reason might be 'surprise' is an emotion that could also be viewed as positive or negative, so maybe hard to express in text.

### 5.4.3 Inter-approaches analysis

The confusion matrices 5.6 and 5.7 show that, with respect to all classes, the performance of the classifier with the personalised model is higher than the global model.

'Neutral' was the class where the classifier tended to assign entries to all the other emotions.

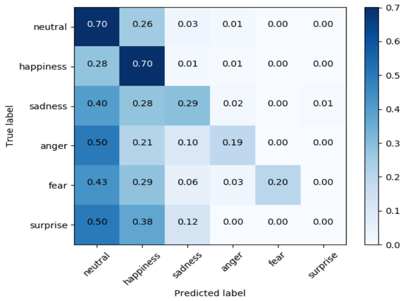


FIGURE 5.6: Confusion matrix for Ekman's personalised model using APP-BOW

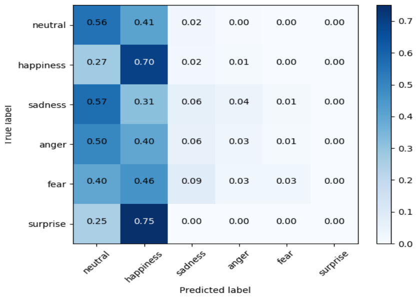


FIGURE 5.7: Confusion matrix for Ekman's global model using Act-BOW

For a deeper inspection of the results, the dis-aggregated data is investigated by generating box plots to visualise the differences in the classifier performance across the participants, to see if some individuals are easier to predict. The median performance is shown with a horizontal line in the box, and the upper and lower quarterlies displayed on the top and the bottom of the box respectively.

Figures 5.8 and 5.9 show that, in both personalised and global, 'happiness' is better predicted when using the combined set of features. In the personalised model, precision is high and with global recall is high *indicating that in contrast to the negative emotions 'happiness' is a universal emotion that may have common features between people.*

Although 'Sadness' is not predicted globally, it is weakly predicted with the personalised model as can be seen in Table 5.2 and the box plot in Figure 5.10 where the boxes are comparatively tall (relevant to 'happy').

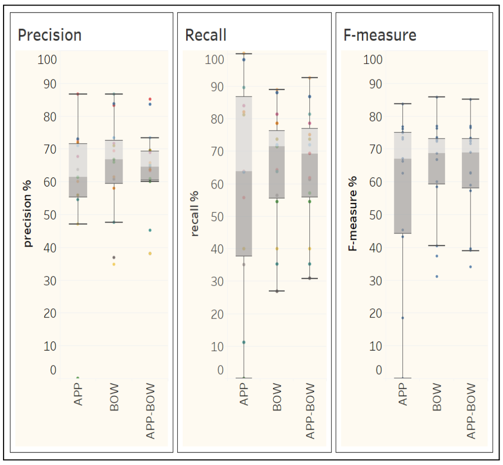


Figure 5.8: Ekman's 'happiness' prediction in the personalised model

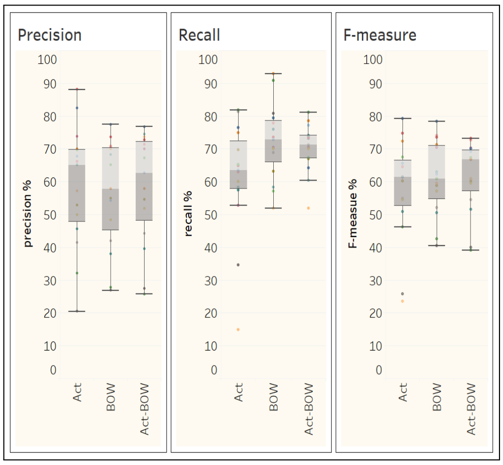


Figure 5.9: Ekman's 'happiness' prediction in the global model

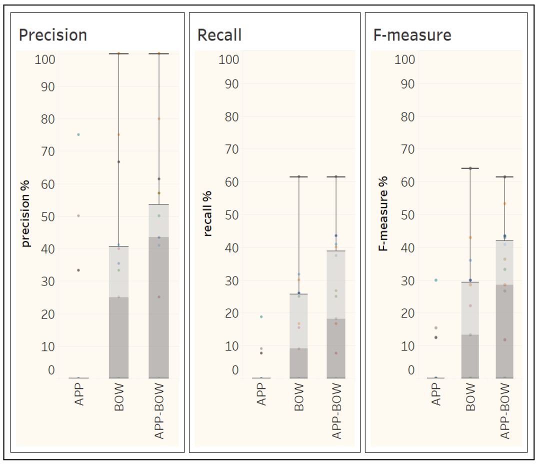


Figure 5.10: Ekman's 'sadness' prediction in the personalised model

This indicates that the classifier behaved very differently with different people's data, it is clear from the median that when adding APP to BOW enhanced prediction of this class, and may indicate that *'sadness' is an emotion that can be determined from activities, people, and places in combination with the language used to describe the situation.*

### 5.4.4 Statistical analysis testing

The previous sections discussed emotion prediction results within different levels. In this section, statistical hypothesis testing is conducted to further assess confidence in the results. Cochran's Q and McNemar's test are applied here to compare the performance of the multiple classification approaches using different feature sets, to check if the models disagree in the same or different ways. Cochran's test is a generalised version of McNemar's test, i.e. only if Cochran's Q test returned statistically significant results, should it be followed by pair wise post hoc analysis using McNemar's test. McNemar's test is a well-known test to analyse the statistical differences in classifier performance, recommended by [194] to compare classifiers when there is both a limited amount of data and it is expensive to train the classifier. The statistic reports on different correct or incorrect predictions between models at the entry level. This test is a type of marginal homogeneity test in the contingency tables that relies on the fact that both classifiers were trained on the same training data and evaluated using the same testing data set.

Statistical significance of the results was examined by comparing the the p-values against $\alpha = 0.05$. Cochran's Q test on the Act; BOW and Act-BOW returned insignificant results with p-value= 0.145 with the global model. So, no further analysis was conducted. With the personalised learning model, running Cochran's Q test on APP; BOW; and APP-BOW returned p-value = 0.004 which showed that there is a significant difference between the three approaches. Further analysis using McNemar's test (Table 5.4) showed that the two approaches APP and BOW are significantly different, meaning that they disagreed in different ways and have different error proportions, as may be expected. Also, APP and APP-BOW are significantly different indicating APP behaved differently when BOW was an additional feature. However, adding APP to BOW did not affect its behaviour; the result was not significant as they disagreed in the same way.

| **Approaches** | **Personalised** |
|---|---|
| APP & BOW | **0.008** |
| BOW & APP-BOW | 0.250 |
| APP & APP-BOW | **0.036** |

Table 5.4: P-values for McNemar's test in Ekman's. Bold values show significance using $p < 0.05$

## 5.5 Circumplex emotion scheme

The Circumplex model, shown in Figure 5.11, is a balanced scheme over eight positive and negative emotions. *Happiness*, *excitement*, *relaxation*, and *calmness* represent positive emotion, while the four negatives are *upset*, *stress*, *tense*, and *boredom*, with *neutral* showing an absence of an emotional state.

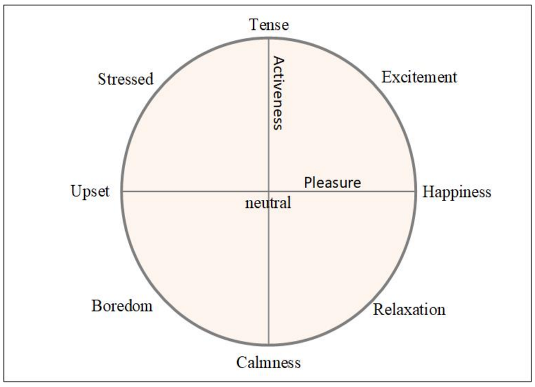


FIGURE 5.11: The Circumplex model

### 5.5.1 Emotion distribution

The full emotions distribution for Circumplex model aggregated over all users is displayed in Figure 5.12 as P0, followed by the distribution in each individual's data. This shows that the two most common emotions were 'neutral' and 'happiness', with positive emotions surpassing the negative ones. Collectively 'Boredom' was expressed at least once by half of the users (an emotion not explicitly reflected in Ekman's scheme) while 'tense', 'stress', 'relaxation' were rarely expressed.

Analysis of the data revealed that the class distribution is imbalanced. As most classification algorithms assume a balanced class distribution, this may provide unfavourable accuracy across the classes. Therefore, to avoid such problem, the *cost-sensitive learning* method was used here. This method works by re-weighting the training instances according to the total cost assigned to each class using a cost matrix, as it is a numerical representation of a misclassification penalty [195, 196]. Other methods are available to balance the data, such as over-/under-sampling; however, the cost-sensitive method was found to be a suitable choice to balance the classes in the Circumplex due to the

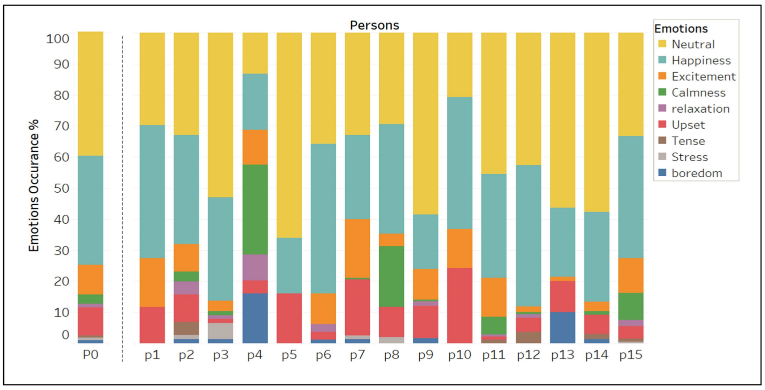


Figure 5.12: The Circumplex emotions distribution

additional structure between the classes. The cost matrix shown in Table 5.5 was generated based on graph distance between the emotions in the Circumplex model shown in Figure 5.11, for example, showing that 'happiness' is three "steps" removed from 'stress'. It should be noted here that the penalty given for misclassifying emotions as 'neutral' was 4 intended to encourage classifications towards concrete emotions.

| | | Predicted | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | neutral | excitement | happiness | relaxation | calmness | boredom | upset | stressed | tense |
| Actual | neutral | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | excitement | 4 | 0 | 1 | 2 | 3 | 4 | 3 | 2 | 1 |
| | happiness | 4 | 1 | 0 | 1 | 2 | 3 | 4 | 3 | 2 |
| | relaxation | 4 | 2 | 1 | 0 | 1 | 2 | 3 | 4 | 3 |
| | calmness | 4 | 3 | 2 | 1 | 0 | 1 | 2 | 3 | 4 |
| | boredom | 4 | 4 | 3 | 2 | 1 | 0 | 1 | 2 | 3 |
| | upset | 4 | 3 | 4 | 3 | 2 | 1 | 0 | 1 | 2 |
| | stress | 4 | 2 | 3 | 4 | 3 | 2 | 1 | 0 | 1 |
| | tense | 4 | 1 | 2 | 3 | 4 | 3 | 2 | 1 | 0 |

Table 5.5: Cost matrix

To see the effect of balancing the data classes, the results for one feature set are presented here in Tables 5.6 and 5.7, that show the personalised prediction results for the imbalanced and balanced model respectively. It is obvious that using the cost sensitive balanced learning enhanced recall for all 'emotional' groups, but not with 'neutral'. The results show the effect of a high penalty for classifying 'neutral', resulting in a small drop in recall for this class offset by increases for all others. So, the balanced model will be used in all further results.

| | **APP-BOW** | | |
|---|---|---|---|
| **Emotion** | **Precision** | **Recall** | **F-measure** |
| neutral | 54.0 | 63.9 | 58.3 |
| happiness | 61.1 | 59.4 | 59.1 |
| excitement | 28.3 | 18.9 | 22.2 |
| calmness | 16.1 | 18.7 | 17.2 |
| relaxation | 7.5 | 4.7 | 5.8 |
| upset | 30.0 | 20.4 | 25.7 |
| tense | 15.0 | 13.3 | 14.1 |
| stress | 20.0 | 5.0 | 8.0 |
| boredom | 15.2 | 13.5 | 13.2 |

Table 5.6: The Circumplex: personalised imbalanced prediction results

| | **APP-BOW** | | |
|---|---|---|---|
| **Emotion** | **Precision** | **Recall** | **F-measure** |
| neutral | **55.2** | 55.5 | 54.2 |
| happiness | 58.1 | **66** | **61.3** |
| excitement | **28.7** | **22** | **23.9** |
| calmness | **17.1** | **21.6** | **18.9** |
| relaxation | 7.5 | 4.7 | 5.8 |
| upset | **31.6** | **22.1** | 24.2 |
| tense | 14.54 | **16.0** | **15.2** |
| stress | **33.3** | **15.0** | **19.4** |
| boredom | 14.5 | 13.5 | **13.9** |

Table 5.7: The Circumplex: personalised balanced prediction results

### 5.5.2 Prediction results analysis

The results of emotion prediction using personalised and global training models are displayed in Table 5.8 and 5.9.

A clear outcome is that *'happiness' is well predicted, and strongly associated with activities*, with a high recall of this emotion with both learning methods, personalised and global.

There was no association between 'tense', 'stress', 'boredom', and 'relaxation' with activities, as indicating *they may be associated with some external circumstances rather than "what" people are doing.* Although they are predicted with the other approaches, the performance was poor. Hence, it is hard to train and evaluate the classifier to predict such emotions without using additional language features (associated/not associated with other features e.g., people and places). Thus, *emotion can't be determined based purely on ones' activities.*

'Excitement', 'calmness', and 'upset' were poorly predicted, precision is better than recall, still not sufficient to evaluate the classifier.

Globally, with activities only 'neutral' and 'happiness' could be predicted. It was also observed that taking entries labelled with 'tense', 'stress' and 'boredom' that could be predicted in the personalised model and putting them in the global model made it impossible for the classifier to predict them. *Predicting such emotions is possible*

| | Act | | | APP | | | BOW | | | Act-BOW | | | APP-BOW | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Emotion** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** |
| neutral | 44.6 | 24.4 | 26.6 | 48.6 | 34.4 | 36.4 | 54.4 | 53.5 | 52.7 | **55.7** | **55.7** | **54.4** | 55.2 | 55.5 | 54.2 |
| happiness | 41.7 | **83.6** | 54.2 | 45.1 | 73.8 | 52.6 | 55.9 | 66.0 | 60.0 | **58.8** | 67.0 | **62.1** | 58.1 | 66.0 | 61.3 |
| excitement | 3.4 | 5.0 | 3.6 | 4.7 | 3.0 | 3.4 | **31.4** | **23.0** | **26.2** | 29.1 | 22.0 | 24.7 | 28.7 | 22.0 | 23.9 |
| calmness | 3.4 | 10.0 | 5.1 | 14.3 | 13.8 | 10.8 | 19.3 | 21.6 | 19.8 | **21.1** | **24.9** | **22.7** | 17.1 | 21.6 | 18.9 |
| relaxation | 0 | 0 | 0 | 0 | 0 | 0 | **9.4** | **4.7** | **6.3** | **9.4** | **4.7** | **6.3** | 7.5 | 4.7 | 5.8 |
| upset | 6.6 | 6.0 | 5.2 | 6.8 | 8.3 | 6.7 | 29.6 | 22.2 | 23.6 | **32.4** | **23.3** | **25.3** | 31.6 | 22.1 | 24.2 |
| tense | 0 | 0 | 0 | 12.5 | 11.72 | 12 | **20** | **20** | **20** | 14.54 | 16 | 15.24 | 14.54 | 16 | 15.2 |
| stress | 0 | 0 | 0 | 0 | 0 | 0 | 33.3 | 15 | 19.4 | 33.3 | 15 | 19.4 | 33.3 | 15 | 19.4 |
| boredom | 0 | 0 | 0 | 0 | 0 | 0 | 11.3 | 11.7 | 11.1 | 14.5 | 13.5 | 13.9 | 14.5 | 13.5 | 13.9 |

TABLE 5.8: Circumplex personalised prediction. Bold values indicate the highest values across approaches

| | Act | | | APP | | | BOW | | | Act-BOW | | | APP-BOW | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Emotion** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** |
| neutral | **58.6** | 22.9 | 30.4 | 53.1 | 27.3 | 33.6 | 50.3 | 38.6 | 41.7 | 52.6 | 39.9 | 44.9 | 51.6 | **45.8** | **47.0** |
| happiness | 38.4 | **95.2** | 53.5 | 39.9 | 90.2 | **54.0** | 41.4 | 68.3 | 48.8 | 44.2 | 68.1 | 46.4 | **44.9** | 64.5 | 50.0 |
| excitement | 0 | 0 | 0 | 0 | 0 | 0 | 17.2 | 17.0 | 9.3 | 16.6 | **17.8** | **10.6** | **17.6** | 8.2 | 9.4 |
| calmness | 0 | 0 | 0 | 0 | 0 | 0 | **24.1** | 3.9 | 6.7 | 21.7 | 4.5 | 7.26 | 22.5 | **9.2** | **13** |
| relaxation | 0 | 0 | 0 | 0 | 0 | 0 | **20.0** | **5.0** | **8.0** | 12.5 | 3.1 | 5.0 | 11.1 | 2.8 | 4.4 |
| upset | 0 | 0 | 0 | 7.4 | 4.93 | 4.0 | 30.0 | 19.8 | 18.8 | **32.1** | **22.5** | **21.5** | 26.3 | 18.8 | 18.5 |
| tense | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| stress | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| boredom | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

TABLE 5.9: Circumplex global prediction. Bold values indicate the highest values across approaches

*using personal circumstances; however, correlations with features are lost when looking at everyone.*

### 5.5.3 Inter-approaches analysis

Similar to Ekman's, with the personalised model, the classifier performed better with respect to all emotions classes as shown in Figure 5.13. Most of the confusion was with 'neutral' class in both personalised and global models. Globally, the confusion is high between 'excitement' and 'happiness' which supports the finding that happiness is *a universal emotion* and maybe associated with similar activities, word expressions and people and places as shown in the confusion matrix in Figure 5.14.

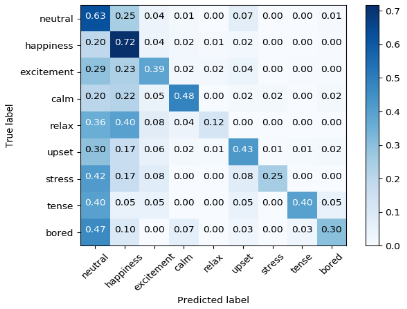


FIGURE 5.13: Confusion matrix for the Circumplex personalised model using BOW-APP

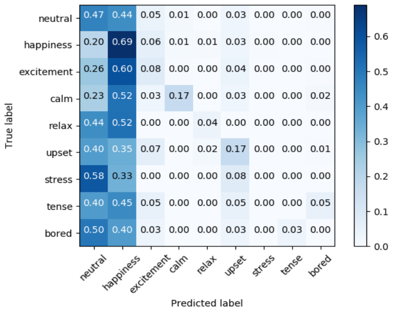


FIGURE 5.14: Confusion matrix for the Circumplex global model using BOW-APP

Figures 5.15 and 5.16 show precision, recall, and F-measure of predicting 'happiness' in both personal and global training. It is clear that *'happiness' is strongly related to some daily activities* as recall was the highest using activities on their own as features for prediction, with global better than personalised. The precision of predicting this emotion is always enhanced when adding new features. Thus, adding people and places may resolves the confusion for the classifier to return entries that are labelled with 'happiness'.

Investigating 'neutral' in Figures 5.17 and 5.18, in both personalised and global learning, there is a clear observation that *when more features were added, the recall was enhanced*, and it reached the best performance with a combination of all features (BOW and APP). This indicates that these features helped the classifier to discriminate between emotional and non-emotional (neutral) entries and returns entries with 'neutral' more accurately.

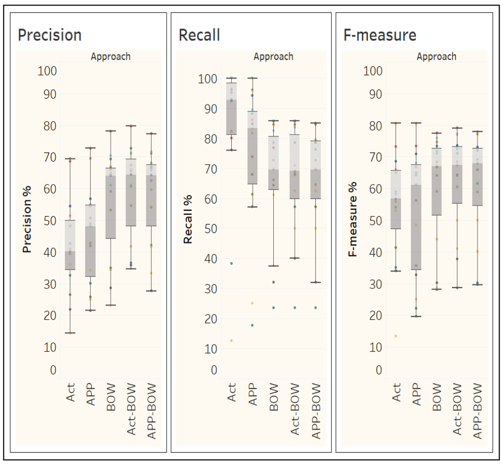


FIGURE 5.15: Circumplex 'happeniess' prediction in the personalised model

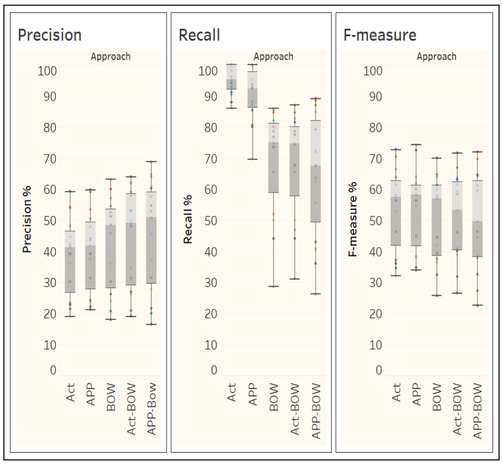


FIGURE 5.16: Circumplex 'happiness' prediction in the global model

Another observation is that the boxes are comparatively larger in the personalised model (particularly with recall) when compared with global learning. This indicates that the classifier behaved very differently across different people's data when training was personalised, while with global training, this variation decreased.

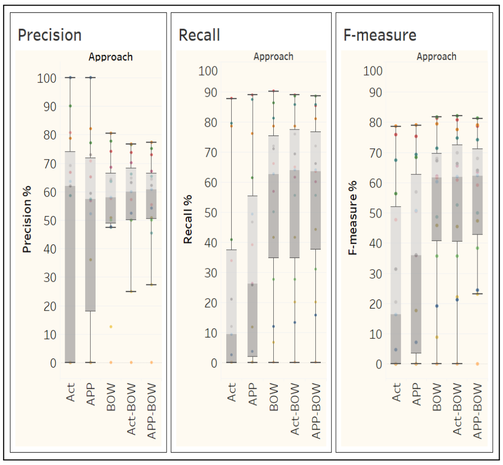


FIGURE 5.17: Circumplex 'neutral' prediction in the personalised model

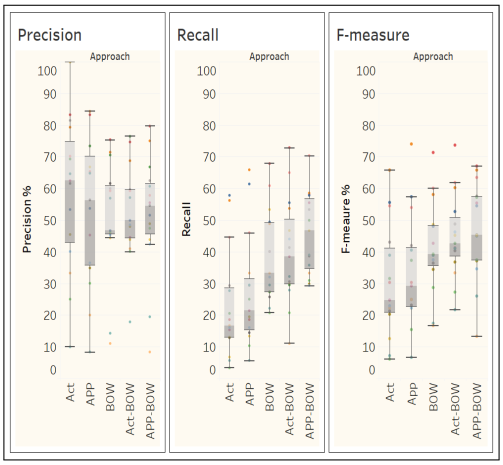


FIGURE 5.18: Circumplex 'neutral' prediction in the global model

### 5.5.4 Statistical analysis testing

As introduced in Section 5.4.4, Cochran's Q and McNemar's tests are conducted to check if there is a statistical difference between the proposed approaches. Cochran's Q test returned significant results on the five approaches: Act; APP; BOW; Act-BOW and APP-BOW with p-values of 0.000 and 0.021 from personalised and global models respectively. Therefore, post hoc analysis was conducted using McNemar's test for both learning methods, Table 5.10 shows the p-values of the test.

| **Approaches** | **Personalised** | **Global** |
|---|---|---|
| Act & BOW | **0.000** | 0.931 |
| APP & BOW | **0.000** | 0.861 |
| BOW & Act-BOW | **0.048** | **0.005** |
| BOW & APP-BOW | 0.444 | **0.000** |
| Act & Act-BOW | **0.000** | 0.300 |
| APP & APP-BOW | **0.000** | **0.018** |

Table 5.10: P-values for McNemar's test in the Circumplex. Bold values show significance using $p < 0.05$

It is clear that with the personalised model, all approaches were significantly different, thus they disagreed with different error proportions, except for BOW and APP-BOW. This indicates that adding people and places may have cancelled the role of activities, meaning that the language used to describe the situation has stronger effect on classification with the personalised model.

Surprisingly, with the global model, the differences between Act (with/without people and places) and BOW were insignificant. Although Table 5.9 showed a clear difference in classification measures, it seems that the entries have similar proportions of disagreement. The difference between BOW and (Act-BOW/APP-BOW) was significant, indicating that adding activities/people and places to BOW made a difference in the classification error rate, thereby causing different behaviour. On the contrary, adding BOW to activities returned insignificant results indicating that *for the global approach activities alone can predict emotions, with little benefit from adding features like BOW.*

## 5.6 The pleasantness and unpleasantness emotion scheme

It is clear from the previous analysis of emotion models that there is some confusion between different types of emotions. Therefore, a much simplified model is investigated here, one that divides the emotional classes into positive emotions (pleasant) and negative emotions (unpleasant).

### 5.6.1 Emotions distribution

Figure 5.19 below shows emotion distribution in the pleasant/unpleasant model. The aggregated data all users is represented in P0, followed with the distribution of emotions in each individual's data. Pleasant emotion class surpassed the unpleasant class for all individuals with neutral/pleasant relatively balanced.

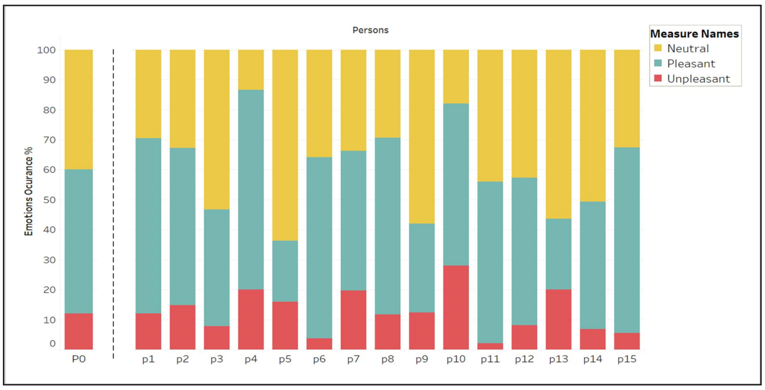


Figure 5.19: The pleasantness/unpleasantness emotions distribution

### 5.6.2 Prediction results analysis

Tables 5.11 and 5.12 below display the personalised and global prediction respectively. 'Pleasant' is the best predicted, 'neutral' is also well-predicted. 'Unpleasant' is poorly predicted and couldn't be predicted globally with activities alone, similar to the previous emotion models, where entries associated with this emotion could not be predicted when trained globally. Another observation is that combining BOW with (activities /people and places) is always better with personalised data.

### 5.6.3 Inter-approaches analysis

The confusion matrices in Figures 5.20 and 5.21 for personalised and global training. The same observation is applied here, with respect to all classes, the personalised learning model returns better classification results than the global model.

Figures 5.22 and 5.23 show the results for predicting 'pleasant' emotions in both personalised and global training.

| | Act | | | APP | | | BOW | | | Act-BOW | | | APP-BOW | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Emotion** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** |
| neutral | 45.4 | 47.2 | 41.8 | 53.1 | 52.7 | 50.3 | **55.6** | 57.7 | 56.3 | 54.5 | 58.5 | 56.1 | 55.4 | **58.7** | **56.7** |
| pleasantness | 51.1 | 67.1 | 57 | 58.2 | 69.2 | 61 | 66 | 68.6 | 66.7 | **67.2** | **67.5** | **66.9** | 66 | 67.1 | 66.1 |
| unpleasantness | 4.44 | 0.6 | 1.1 | 28.3 | 9.4 | 11 | 49.3 | 28 | 32.6 | **49.1** | **29.1** | **34.1** | 43.4 | 27.2 | 32 |

Table 5.11: Pleasantness/unpleasantness personalised prediction. Bold values indicate the highest values across approaches

| | Act | | | APP | | | Bow | | | Act-Bow | | | APP-BOW | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Emotion** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** |
| neutral | **53.8** | 52.8 | 51.2 | 50.32 | **65.2** | **54.9** | 51.7 | 44 | 45.6 | 51.7 | 46.5 | 47.9 | 51.7 | 50.76 | 49.9 |
| pleasantness | 58.3 | 73.7 | **63** | 59.7 | 59.9 | 57.7 | 57.2 | **73** | 62 | 58.2 | 71.4 | 62.2 | **59.8** | 69.1 | 62.5 |
| unpleasantness | 0 | 0 | 0 | 2.1 | 1.1 | 1.4 | **40.8** | **17.6** | **19.5** | 36.8 | 16.9 | 18.1 | 39.4 | 16.5 | 18.1 |

Table 5.12: Pleasantness/unpleasantness global prediction. Bold values indicate the highest values across approaches

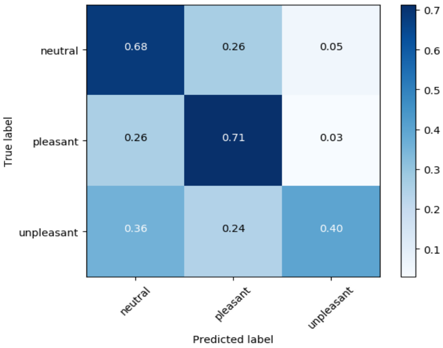


FIGURE 5.20: Confusion matrix for pleasantness/unpleasantness personalised model using APP-BOW

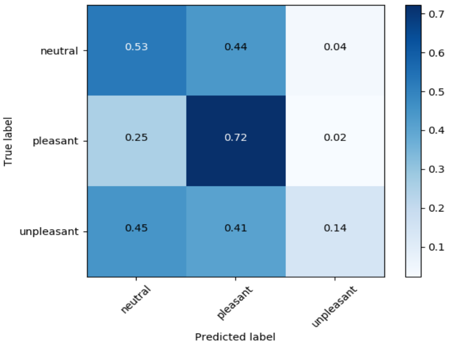


FIGURE 5.21: Confusion matrix for pleasantness/unpleasantness global model using APP-BOW

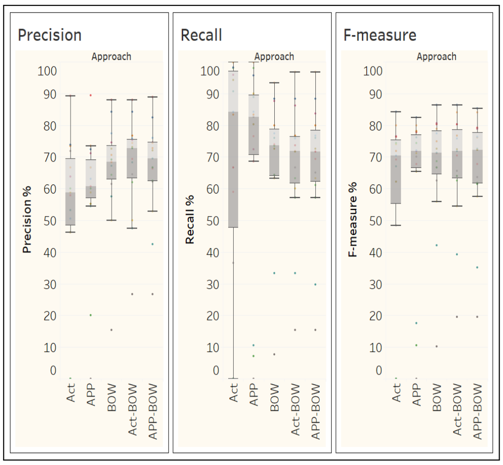


FIGURE 5.22: Pleasantness prediction in the personalised model

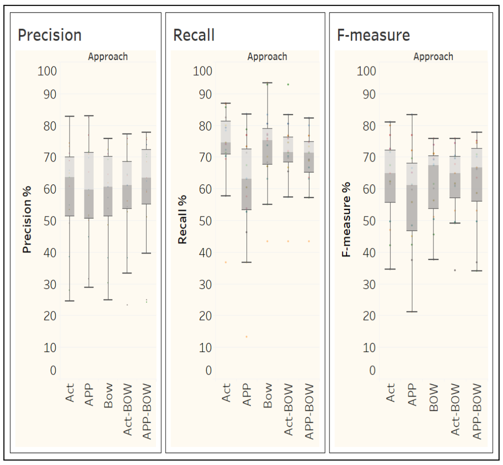


FIGURE 5.23: Pleasantness prediction in the global model

It could be concluded that with global training, the prediction results for all participants are close to each other (accumulated), while they vary greatly with the personalised model (sparse). This makes sense because people are different; different in the way they express themselves and in the frequencies of expressing their emotions. This can be seen clearly with the recall using activities for prediction where the diversity is high with personalised training, while the results are very high and close to each other with global training. Another observation is that associating people and places with activities enhanced prediction with the personalised model while decreasing performance with the global model.

### 5.6.4 Statistical analysis testing

As explained in Section 5.4.4, Cochran's Q and McNemar's tests were conducted to check if there is a statistical difference between the proposed approaches. Cochran's Q test returned significant results on the five approaches: Act; APP; BOW; Act-BOW; and APP-BOW with p-value of 0.000 for personalised models, while with global models the results were insignificant with p-value= 0.073. Table 5.13 shows the p-values of the McNemar's test for personalised training model. As expected, there is a significant difference between (Act/APP) and BOW, as these are completely different feature sets that classifier used differently when interpreting the data. Combining BOW to activities/people and places returned significant results meaning that it disagreed in a different way. Also, adding activities to BOW caused a change in behaviour; however, adding people and places resulted in an insignificant difference.

| **Approaches** | **Personalised** |
|---|---|
| Act & BOW | **0.000** |
| APP & BOW | **0.000** |
| BOW & Act-BOW | **0.045** |
| BOW & APP-BOW | 0.056 |
| Act & Act-BOW | **0.002** |
| APP & APP-BOW | **0.003** |

Table 5.13: P-values for McNemar's test in pleasantness/unpleasantness model. Bold values show significance using $p < 0.05$

## 5.7 The neutral and emotional scheme

This model attempts to investigate a binary classification model to filter the data and differentiate between the emotional and non-emotional (neutral) entries, to investigate if this simpler problem is more tractable for a classifier.

### 5.7.1 Emotions distribution

Figure 5.24 shows a balanced distribution between 'emotional' and 'neutral' classes aggregated from all-users data as P0, followed by the distribution in each individual's data.

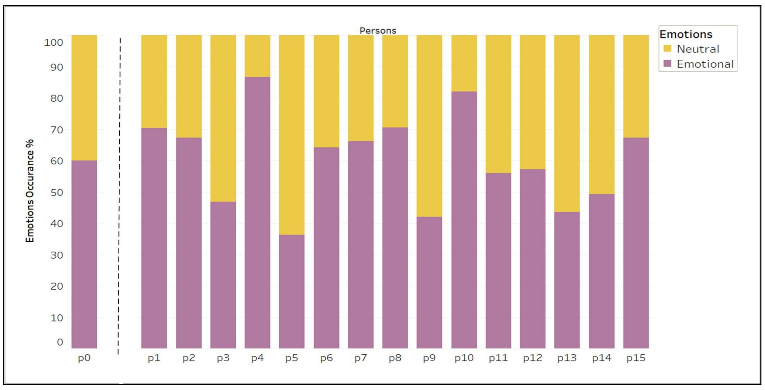


Figure 5.24: Emotional and neutral distribution

### 5.7.2 Prediction results analysis

Table 5.14 shows the average results for the global training model. This experiment aimed to investigate the classifier performance in differentiating between emotional and neutral entries and to explore how adding 'people' as feature to activities affects classification.

| | **Act** | | | **Act-Ppl** | | |
|---|---|---|---|---|---|---|
| **Emotion** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** |
| neutral | **58.7** | 22.1 | 30.7 | 56.8 | **34.9** | **40.9** |
| emotional | 63.6 | **89.8** | **73.7** | **66.1** | 83.34 | 72.62 |

Table 5.14: Emotional/neutral global prediction

The main findings from this model are: performance is very high when predicting emotional class using activities and recall decreased slightly when 'people' was added while it enhanced 'neutral' prediction. This (together with Tables 5.3, 5.9, 5.12) shows that *activities indicate the presence of emotions but do not differentiate between individual emotions*, i.e. it is easier to predict all the instances where people are emotional but harder to pick which emotion they are experiencing.

### 5.7.3 Inter-approaches analysis

Figures 5.25 and 5.26 display the box plots for predicting 'emotional' and 'neutral' classes respectively. They support the two observations from Table 5.14. Firstly, activities (or activities with people) are good predictors for emotional entries. Secondly, adding 'people' to activities caused a slight enhancement with 'neutral' recall and a slight decrease with 'emotional' recall.

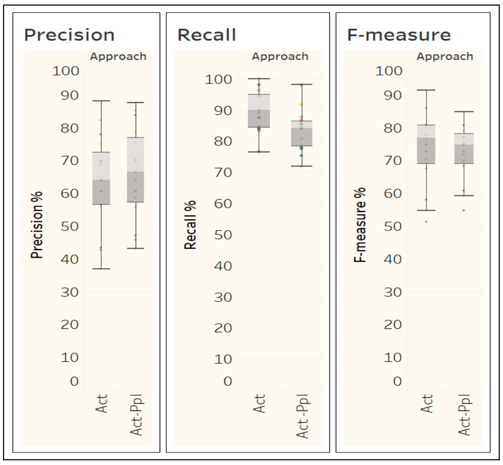


Figure 5.25: 'Emotional' prediction results

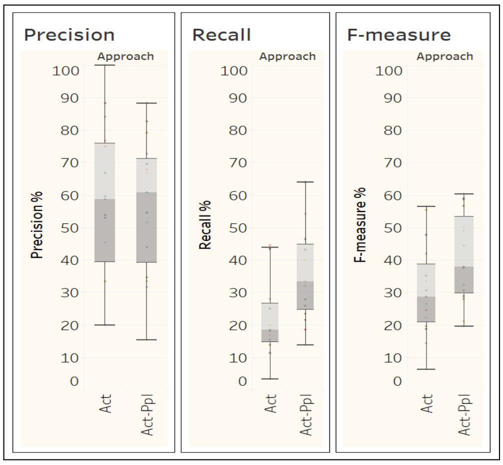


Figure 5.26: 'Neutral' prediction results

A closer look at the results when adding people as features to activities, slightly enhanced prediction with some participants and decreased performance with others. Figure 5.27 shows the weighted average for precision, recall, and F-measure for each participant. It is clear that 'people' as a feature when added to activities enhanced classification with around 60% of the participants, especially with precision, which increased the discrimination ability of the classifier. It has been shown here that this feature affected classification i.e. there is a correlation between mentioned people and absence of emotion. However, there might be some lurking factors, from the machine learning perspective, such as the number of names in one's data set, repetition, association with the same activity, some using different nicknames for a person (e.g., Moody or Mohammed), etc.

### 5.7.4 An intra-person analysis

This section explores one participant's data and investigates the combination of the different features to train a binary (neutral and emotional) classifier. This data set was found to be suitable for machine learning testing; the size of the data is acceptable (283 entries), the length of entries average is around 11 words, more routine entries (repetition) in addition to fewer out-of-routine ones.

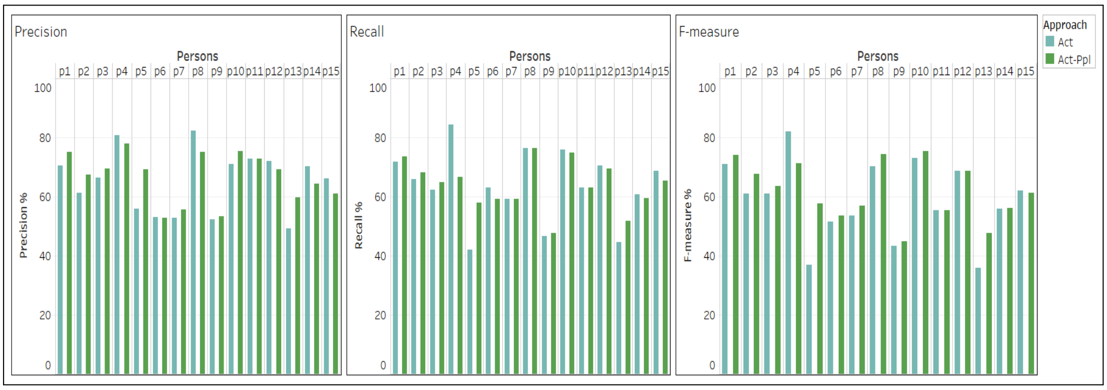


Figure 5.27: Weighted average of the precision, recall, F-measure for each person for both approaches Act and Act-Ppl in the emotional/neutral global model

Table 5.15 shows that recall of 'emotional' detection is very high in personalised training, while 'neutral' was poorly detected. With global training, as shown in Table 5.16, 'neutral' recall was enhanced significantly compared with personalised training. The highest recall for 'emotional' was with activities only indicating that this is a kind of a person, where activities determine emotions even more than the language used. The highest recall for 'neutral' was globally using a combination of all features (activities, people, places, and BOW).

| | **Neutral** | | | **Emotional** | | |
|---|---|---|---|---|---|---|
| **Approach** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** |
| Act | 50 | 1.2 | 2.4 | 70.6 | **99.5** | **82.6** |
| Act-Ppl | **60** | 14.5 | 23.2 | 72.8 | 96 | 82.8 |
| Act-Loc | 33.3 | 4.8 | 8.4 | 70.6 | 96 | 81.4 |
| BOW | 53.4 | 47 | 50 | 78.8 | 82.8 | 80.8 |
| BOW-Ppl | 54.4 | 44.6 | 49 | 78.4 | 84.3 | 81.3 |
| BOW-Loc | 54.2 | 47 | **50.3** | **78.9** | 83.3 | 81.1 |
| BOW-Ppl-Loc | 55.1 | **45.8** | 50 | 78.8 | 84.3 | 81.5 |
| BOW-Act | 52.1 | 45.8 | 48.7 | 78.4 | 82.3 | 80.3 |
| BOW-Act-Ppl | 52.1 | 44.6 | 48.1 | 78.1 | 82.8 | 80.4 |
| BOW-Act-Loc | 52.8 | **45.8** | 49 | 78.5 | 82.8 | 80.6 |
| BOW-Act-Ppl-Loc | 52.1 | 44.6 | 48.1 | 78.1 | 82.8 | 80.4 |

Table 5.15: Prediction results for personalised training for one participant

| | Neutral | | | Emotional | | |
|---|---|---|---|---|---|---|
| **Approaches** | **Precision** | **Recall** | **F-measure** | **Precision** | **Recall** | **F-measure** |
| Act | 52.9 | 43.4 | 47.7 | 77.9 | **83.8** | **80.8** |
| Act-Ppl | 54.6 | 63.9 | 58.9 | 83.7 | 77.8 | 80.6 |
| Act-Loc | **55.1** | 65.1 | **59.7** | 84.2 | 77.8 | **80.8** |
| BOW | 45.7 | 77.1 | 57.4 | 86.5 | 61.6 | 72 |
| BOW-Ppl | 45.5 | 79.5 | 57.9 | 87.5 | 60.1 | 71.3 |
| BOW-Loc | 46.4 | 77.1 | 57.9 | 86.7 | 62.6 | 72.7 |
| BOW-Ppl-loc | 45.6 | 80.7 | 58.3 | 88.1 | 59.6 | 71.1 |
| BOW-Act | 43.2 | 65.1 | 51.9 | 81.4 | 64.1 | 71.8 |
| BOW-Act-Ppl | 44.3 | 79.5 | 56.9 | 87.1 | 58.1 | 69.7 |
| BOW-Act-Loc | 46.1 | 78.3 | 58 | 87.1 | 61.6 | 72.2 |
| BOW-Act-Ppl-Loc | 43.7 | **83.1** | 57.3 | **88.6** | 55.1 | 67.9 |

Table 5.16: Prediction results for global training for one participant

The results in Table 5.17 show the effect of adding places (Loc) as a feature, most of times it enhanced prediction slightly. Sometimes, it returned the best classification results (e.g., Act-Loc when predicting 'neutral' and 'emotional' in the global model).

| **Approaches** | Personalised | | Global | |
|---|---|---|---|---|
| | **Neutral** | **Emotional** | **Neutral** | **Emotional** |
| Act vs. Act-Loc | ↑ | ↓ | ↑ | - |
| BOW vs. BOW-Loc | ↑ | ↑ | ↑ | ↑ |
| BOW-Ppl vs. BOW-Ppl-Loc | ↑ | ↑ | ↑ | ↓ |
| BOW-Act vs. BOW-Act-Loc | ↑ | ↑ | ↑ | ↑ |
| BOW-Act-Ppl vs. BOW-Act-Ppl-Loc | - | - | ↑ | ↓ |

Table 5.17: The effect of adding places (Loc) as a feature

## 5.8 Discussion

One obvious question is, whether these results are dependent on the classifier used? To address this, an exploratory experiment testing other classification algorithm was conducted using Weka. The activity vectors following Section 3.8 were used as features for prediction, and ten-fold cross-validation was applied to the labelled entries for the training and testing of the different classifiers. The tested algorithms were neural networks NN (Multi-layers Perceptron), NB, DT and SVM. Table 5.18 displays the results and shows the accuracy for emotion prediction. The columns show the different algorithms that were tested and the rows show the emotion models. It is clear from this table that NB returned the lowest results while NN, DT and SVM returned close results with SVM slightly higher. NN was not a suitable choice here because they take a long time to develop and train; more importantly, they need too much data to train.

| | Neural Networks | Naive Bayes | Decision Tree | Support Vector Machine |
|---|---|---|---|---|
| Circumplex | 52.4 | 40.0 | 52.6 | 52.6 |
| Ekman | 56.5 | 44.3 | 55.8 | 57.1 |
| Pleasantness/ Unpleasantness | 57.6 | 53.7 | 57.7 | 58.6 |
| Emotional/ Neutral | 64.28 | 62.95 | 63.6 | 66.92 |

Table 5.18: Classification accuracy for different algorithms on different emotion models

The emotion prediction in this chapter is based on the activities vectors of a small data set. Thus, it is not only about the words in each entry, but it is also about the activity this entry is referring to, unlike the classification of activities in the previous two chapters where words represented the main feature for classification and pre-trained NN model, which was trained with 1.2 million vocabularies, was a suitable choice. Therefore, SVM was preferred here as it outperformed the other algorithms.

In all the three emotion models (Ekman's, the Circumplex, and the pleasantness/unpleasantness), prediction using the personalised training is always better than the global. This suggests that the triggers for emotional responses and/or their expression in diaries varies considerably and that an emotion classifier should be more accurate when trained with personal data. However, this requires collecting a significant volume of data per individual over a long time. This finding was also confirmed statistically, as Cochran's Q test results were always significant when conducted for the personalised model investigation.

In general, analysing this data set showed that 'neutral' and 'happiness (pleasantness)' are the best-predicted emotions. One possible reason for this is that they represented the classes with the highest number of entries, while the negative classes have a lower number of entries. This may also be due to *cultural norms* that influence responses with emotion self-reporting, as highlighted by Scollon et al. [197]. For example, if there is a cultural norm that feeling negative emotions is undesirable, some people may not report their emotions as 'sadness'. Another reason might be *social desirability* [198] when reporting feelings or attitude, where some "defensive" people, for example, may find it difficult to report such negative emotions. Although this problem was solved to some extent with the Circumplex model, where the data was balanced by re-weighting classes, prediction of the negative classes was still poor, whether with personalised or global training.

One notable finding is the extent to which daily activities represented as features vectors (associated or not with people and places) could predict emotions. This returned results that were close to those with BOW and sometimes higher for certain emotions (e.g., ‘happiness’ in global training with all models). Interestingly, statistically testing the results applying McNemar’s test showed significant differences between the BOW and Act/APP (with personalised training) meaning that they had different proportions of errors. Nevertheless, they both showed an ability to predict emotions properly.

For all emotional models, the BOW approach performed acceptably. However, the usage of a small set of discriminating features (i.e. activities, people, and places) slightly enhanced the performance. This can particularly be seen with the underrepresented classes such as ‘sadness’, ‘anger’, ‘fear’ in Ekman’s, and ‘tense’, ‘stress’, ‘boredom’, ‘relaxation’ in the Circumplex, and unpleasant classes. This finding is consistent with the statistical testing results. As McNemar’s test showed, adding activities to BOW changed the classifier’s behaviour. The opposite was also true where adding BOW to Act/APP showed significant differences between the approaches.

There are three main findings from the global training. Firstly, the highest recall in predicting ‘happiness’ was reached with global training for all emotional models, indicating that this is evident and expressed by all emotions where features from different individuals may overlap and strengthen the learning process. Secondly, no negative emotions could be predicted based on any activities, which suggests that there is a lack of commonality across people and indicates that these emotions cannot be determined based only on daily activities. Perhaps, unexpectedly, it is not just what someone is doing that determines her/his negative emotions. Thirdly, the effect of global training can clearly be seen in the Circumplex model. Interestingly, some emotions like ‘tense’, ‘stress’, and ‘boredom’ could not be predicted, although, they were poorly predicted with personalised training. This suggests that it is possible to predict some people are ‘tense’ (for example) based on their personal circumstances, but not with global training features.

Looking at the comparison between different models to group emotions showed some interesting outcomes. For example, predicting ‘pleasantness’ emotion using global training based only on activity as a feature showed an increase in F-measure associated with the decrease in the number of groups (see Table 5.19), indicating that the prediction of this emotion is enhanced when using more features (i.e., activities) from other emotional groups.

| | Happiness/pleasantness/ emotional | | |
|---|---|---|---|
| **Model** | **Precision** | **Recall** | **F-measure** |
| Circumplex | 38.4 | 95.2 | 53.5 |
| Ekman's | 58.5 | 61.3 | 57.5 |
| Pleasantness/unpleasantness | 58.3 | 73.7 | 63.0 |
| Emotional/neutral | 63.6 | 89.8 | 73.7 |

Table 5.19: A comparison between different grouping of emotions

It can be concluded from Tables 5.3, 5.9, 5.12 and 5.14 that *activities can better indicate the presence of emotions rather than to differentiating between individual emotions.* That is, it is easier to predict instances where people are emotional but harder to pick which emotion they are experiencing. Although the performance can be improved by conflating classes, the ability to identify/manage conditions relating to specific emotions is lost, which may affect the application. For example, if a medical practitioner wants to predict a specific emotion such as anger, or wanted to treat a condition like depression, he/she would use emotion models with variety in classes but accept the risk of low performance.

The data used in this research was collected from healthy people (as far as we know). This may explain why it was insufficient to predict people with negative emotions, such as stress and anxiety, that practitioners usually track [199, 200]. However," happiness" is an emotion that can accurately be predicted with global and personal training when using daily activities as features, which indicates that people may have some similarities in which activities are associated with "happiness". Such work may be very useful in line with the latest researches in psychology that are moving more towards studying positive emotions, especially "happiness" [201, 202].

Examining different schemes for measuring emotion raised the possibility that the difficulty of predicting using Ekman's scheme may be associated with its development, where it demonstrates that individuals have universal basic emotions that are shared across cultures are recognisable through facial expressions [203]. Therefore, it is oversimplified where positive emotions can only be referred to by 'happiness'. Accordingly, this scheme does not have enough structure; emotions are "discrete". In contrast, the Circumplex model was originally created by taking account of phrases or words that people use to describe their emotions and placing them into a circular order [82]. Therefore, emotions are distributed in a two-dimensional circular space, in a way that shows some spatial relationship between emotions, for example, ‘happiness’ and ‘upset’ describe opposite emotions, so they are placed opposite to each other, while ‘relaxation’ and ‘calmness’ describe similar emotions, so they are placed closer to each other.

This structure of the Circumplex model motivated the use of the cost-sensitive approach for balanced learning to address the underrepresentation of negative emotions. As a result, the prediction of almost all emotional classes was enhanced by using the personalised model. Also, it showed a significant increase in recall of ‘happiness’ in the global model. However, it was associated with some decrease in the recall for predicting ‘neutral’, but this is not the focus here.

## 5.9 Summary

In this chapter, a multilevel analysis was carried out on four emotion models: Ekman’s; the Circumplex; pleasantness/unpleasantness and emotional/neutral.

Each emotion model was investigated, starting with an analysis of the distribution of emotion classes. This was followed by an aggregated analysis of personalised and global learning models, along with a deeper inspection of the dis-aggregated results. Finally, statistical significance testing was performed to increase confidence in the results. The discussion at the end of this chapter, which provided a detailed analysis of the findings, showed that the findings are consistent with research in the field of psychology that suggests that emotion correlates with various daily activities, people, and places.

# Chapter 6

# Conclusions, limitations and future work

Automating the process of emotion detection from self-reported diaries is extremely valuable, since it brings new insight to the study of individuals' situation (e.g. activity) and environment (e.g. people and places).The aim of the research in this thesis was towards predicting emotion based on people's reports on their daily life or current situation. Activities, people, and places were used to infer emotion (as a dependent factor). The association of these factors with emotion is well-studied in psychology, which motivated interest from a computer science point of view.

Digital textual dairies are a rich resource of people's involvement in activities, relationships, engagement and attachment to places, as well as associated emotions, and were used as the data source to investigate the ability of the machine to understand and analyse this data.

This work was built iteratively, starting by proposing a framework for classifying diary entries in terms of personal activities by combining unsupervised (clustering) and supervised (classification) machine learning techniques. After successfully classifying the entries into their relevant activities, information extraction techniques were used to detect people and places from the text. Emotion was then predicted based on these three features using personalised and global learning strategies, under different emotion schemes, including Ekman's six basic emotions, the Circumplex, pleasantness/unpleasantness and emotional/neutral.

The availability of many digital applications used for data collection in diary methods made it easy to have big data about everyday life; however, there is still a need to improve the analysis process and reduce the need for human intervention. Therefore, this research has investigated possible ways to automate the process of analysis, aiming to enhance analysis further and make results more representative, whether with identifying activities, people and places, or combining these features to predict emotion, unlike the available methods for predicting emotions that mostly focus of the relevant dictionary-based/key-words. The research in this thesis supports improving practitioners' and researchers' efficiency in many important fields such as behavioural science, social science and psychology, and contribute to the development of personalised systems.

Much of the research in computer science to predict activities was conducted for different purposes other than emotion prediction, e.g., travel recommendation systems. These are based on the language and traditional machine learning algorithms. Little research relates activities to emotions using n-gram BoW, and suffers from too much human intervention and the problem of a small training data set. The proposed framework for activity classification provided a novel and efficient solution in response to the challenges of sparsity and lack of training data. Firstly, a transfer learning approach exploits a publicly available and easy to obtain data set similar to diary data (tweets). Secondly, the centroid-based clustering step added flexibility and efficiency to the classification approach by taking advantage of proximity to deal with new similar human-generated data with less training, i.e., a kind of self-trained clusters.

Much research has been conducted on emotion prediction, using different features (e.g. linguistic), and pursuing different goals (e.g., recommendation systems). To the best of our knowledge, the research in this thesis is the first to investigate combining the distinct features that describe events: activities, people and places, to predict emotion for the purpose of providing technical support and improving practitioners' work.

## 6.1 Conclusions

This research has shown that using a small set of automatically detected domain-specific features together represent the individual's current situation, could successfully predict emotions. Moreover, adding activities as a feature to the language feature enhanced the classifier's performance in some cases. However, the performance varied considerably *between emotions* (e.g. 'happiness' is better predicted than others), *between participants* (i.e. some participants were predicted better than others) and *between learning models* (i.e. classification using the personalised model was typically better than the global model). This research has also led to the following conclusions:

- The proposed model for activity categorisation successfully categorised high-level daily activities in a semi-supervised manner, and has largely overcome the challenges of the sparsity of the diary entries and the lack of training data by utilising a pre-trained model on similar, but not identical, inexpensive publicly available data (tweets).
- The investigation of the different approaches to handle trivial diary entries in the activity classification task has emphasised the importance of the decision to include 'none' as a distinct class or to restrict classification to activities only. As it was tempting to classify all entries into relevant activities (i.e. exclude 'none'), the results suggested that including 'none' was a rational decision.
- Regardless of which emotion model is utilised, this thesis has shown that it would be very hard to predict unpleasantness or negative emotions using global training based on activities only; however, combining it with the language features may offer a slight improvement. The performance in predicting 'happiness/pleasantness' may enhance with the decrease in the number of classes.
- An emotion classifier should be more accurate when trained with personal data. In all the different emotion models utilised in this research, prediction using the personalised training was always better than the global. However, this requires collecting a significant volume of data per individual over a long time.
- In general, understanding emotions and predicting them accurately is not an easy task because of the complex interaction between all factors that affect emotions (e.g. illness and mood swings), which can obscure the emotions a person is experiencing. However, machine learning (statistically and mathematically) provides a partial view by automatically exploiting some hidden patterns. The best method will depend on the use case or end application; however, the results of the research in this thesis gives insight into this trade off. For example, if the interest is to get

a high recall rate for predicting people in an emotional state, then using activities with the global training would be the suitable approach for a binary classifier. While if the aim was to get a high precision rate, then personalised training with additional language features is recommended.

In summary, the research in this thesis has looked at people's situations and their environment/circumstances from their perspectives and the personal words they recorded about themselves. Applying the available technologies such as machine learning, information extraction, neural network and choosing the most suitable algorithms and learning strategies created new insights. They demonstrated that automating textual diary analysis would provide a clear image of the authors, predict their emotions, and maybe recommend some new association of factors that can affect their emotions, especially with positive emotion.

## 6.2 Limitations and future work

The emotion prediction model was evaluated with a small-scale user population. Although results were statistically validated and the feasibility of such work was proved, it cannot be guaranteed here that the results will generalise. Therefore, a large-scale investigation remains as future work.
The major application of this work is in clinical fields. So, in an opposite scenario, if we are treating people with depression and the data did not include any positive emotion, a similar problem would occur, i.e., it would be hard to predict an emotion such as "happiness" as it would be underrepresented. It is very difficult to get a representative data set that covers all people in all circumstances as asking people to keep diaries is a common approach in psychology to mental health, but the number of general people who keep diaries is quite low, so there is a kind of mismatch between getting a large number of data and getting a relevant data. In the current situation with the available data set, we showed that the results in this thesis are reliable as they were statistically tested and validated for the different emotion models and learning strategies.

A challenge in studying emotions as experienced in the various types of situations that people encounter in their daily life is the need for frequent and repeated collection of data. As this is a human-generated diary, some challenges using this data were related to the quality of the collected data, where the contents were incomplete and fragmented. A more specialised diary application could be used in future work to collect a larger, more consistent data set, or rather than asking people to record their diary entries purely for the purpose of research, the experiment could be embedded within other features that the participants find useful, thus providing more natural and meaningful data (as was carried out in [204]).

As the work in this thesis looked at only textual diaries, the combination of textual diary data and mobile sensing (as well as social media data) may allow the collection of a large amount of data. As the current study describes how people's emotions vary with different situations, having more relevant data (e.g. heart rate) may help in further exploration of emotions and emotion inference, but it would also provide a challenge in collecting data for the most relevant time.

Evidence was found that people experience differences in emotions when doing different activities, are with different people or are in different places. As research continues to identify the factors of a situation that are psychologically meaningful, future work will be able to investigate the interactive effects of more various situational factors on people's emotions (e.g. objects).

In addition to the features proposed in this research, other features could be further investigated, such as 'personality'; indeed, some studies have already shown that personality traits help to explain the individual differences in the types of places visited [205]. It is fundamental to consider that although activities, people, and places affect an individual's emotions, there are other factors within the person that can affect their emotion at a particular time. As emotions can be influenced by exclusive personal and behavioral factors [44, 206], emotions for an individual participant can change for the same event for unreported reasons.

Chapter 5 has shown that the grouping or the granularity of emotions is a key factor in classification, so it would be interesting to look at grouping within features, for example, grouping places (e.g. indoor and outdoor) or grouping people by relationship types (e.g., family, friends, professional).

The correlations between specific activities and emotions could also be investigated, for example, socialisation could be indicated by a social activity (e.g. visiting my friend), since studies in psychology show that people who socialise more are happier than those who don't and that being with someone is associated with 'happiness'.

The current results motivate further research of within-person variation in emotions, which can reveal interesting patterns that may be masked at mean levels. Moreover, heterogeneity among people can be an important aspect, which can be emphasised by implementing parameters for each participant and demonstrate how the performance level of the utilised model changes accordingly.

As this research showed that the personalised models perform better than the global models when predicting emotions, although collecting a large set of personalised data is a difficult task, it is recommended to integrate the advantages from both models to overcome their weaknesses by building a hybrid model that gradually incorporates personal data with the global model, which was instantiated from others' training data.

This research looked at high-level activities, including work/study, social/family, food/-drink, leisure and essentials. As has been shown, the nature of the groups played an important role in classification. Therefore, the definition of activity categories can be questioned, and a more precise definition of the required groups may help with classification. Furthermore, individuals might perceive activities differently, for example as some people might consider "baking" as 'leisure' others whilst might consider it as 'essentials'. However, in this work, based on the word embedding vector space model used, 'baking' was classified as 'food/drink'. So, more precise definitions and different categories might result in a more accurate outcome. These may be amended according to the relevant application.

# Bibliography

[1] Ed Diener, Stuti Thapa, and Louis Tay. Positive emotions at work. *Annual Review of Organizational Psychology and Organizational Behavior*, 7:451–477, 2020.

[2] Anna Dluzewska Rowe, Julie Fitness, and Leigh Norma Wood. University student and lecturer perceptions of positive emotions in learning. *International Journal of Qualitative Studies in Education*, 28(1):1–20, 2015.

[3] Ilias O Pappas, Panos E Kourouthanassis, Michail N Giannakos, and Vassilios Chrissikopoulos. Shiny happy people buying: the role of emotions on personalized e-shopping. *Electronic Markets*, 24(3):193–206, 2014.

[4] Anthony D Ong, Lizbeth Benson, Alex J Zautra, and Nilam Ram. Emodiversity and biomarkers of inflammation. *Emotion*, 18(1):3, 2018.

[5] Beatrice Bortolato, Thomas N Hyphantis, Sara Valpione, Giulia Perini, Michael Maes, Gerwyn Morris, Marta Kubera, Cristiano A Köhler, Brisa S Fernandes, Brendon Stubbs, et al. Depression in cancer: the many biobehavioral pathways driving tumor progression. *Cancer treatment reviews*, 52:58–70, 2017.

[6] Alan Carr. *Positive psychology: The science of happiness and human strengths.* Routledge, 2013.

[7] Kari Leibowitz and Joar Vittersø. Winter is coming: Wintertime mindset and wellbeing in norway. *International Journal of Wellbeing*, 10(4), 2020.

[8] Aurobind V Iyer, Viral Pasad, Smita R Sankhe, and Karan Prajapati. Emotion based mood enhancing music recommendation. In *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 1573–1577. IEEE, 2017.

[9] Felice N Jacka. Global and epidemiological perspectives on diet and mood. In *The Gut-Brain Axis*, pages 141–158. Elsevier, 2016.

[10] Katherine T Baum, Anjali Desai, Julie Field, Lauren E Miller, Joseph Rausch, and Dean W Beebe. Sleep restriction worsens mood and emotion regulation in adolescents. *Journal of Child Psychology and Psychiatry*, 55(2):180–190, 2014.

[11] Sophie von Stumm. Feeling low, thinking slow? associations between situational cues, mood and cognitive function. *Cognition and Emotion*, 32(8):1545–1558, 2018.

[12] Paula M Niedenthal and François Ric. *Psychology of emotion*. Psychology Press, 2017.

[13] Damodar Suar, Amrit Kumar Jha, Sitanshu Sekhar Das, Priya Alat, and Pooja Patnaik. What do millennials think of their past, present, and future happiness, and where does their happiness reside? *Journal of Constructivist Psychology*, pages 1–17, 2020.

[14] Marta Miret, José Luis Ayuso-Mateos, Jose Sanchez-Moreno, and Eduard Vieta. Depressive disorders and suicide: epidemiology, risk factors, and burden. *Neuroscience & Biobehavioral Reviews*, 37(10):2372–2374, 2013.

[15] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

[16] Yu Yin, Mohsen Nabian, and Sarah Ostadabbas. Facial expression and peripheral physiology fusion to decode individualized affective experience. In *Workshop on Artificial Intelligence in Affective Computing*, pages 10–26. PMLR, 2020.

[17] Pooja Megha Nagar, Oksana Caivano, and Victoria Talwar. The role of empathy in children's costly prosocial lie-telling behaviour. *Infant and Child Development*, 29(4):e2179, 2020.

[18] Erik Cambria. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107, 2016.

[19] Dennis Becker, Ward van Breda, Burkhardt Funk, Mark Hoogendoorn, Jeroen Ruwaard, and Heleen Riper. Predictive modeling in e-mental health: A common language framework. *Internet interventions*, 12:57–67, 2018.

[20] Anja Thieme, Danielle Belgrave, and Gavin Doherty. Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(5):1–53, 2020.

[21] M Shamim Hossain, Ghulam Muhammad, Mohammed F Alhamid, Biao Song, and Khaled Al-Mutib. Audio-visual emotion recognition using big data towards 5g. *Mobile Networks and Applications*, 21(5):753–763, 2016.

[22] Kevin B Meehan, Chiara De Panfilis, Nicole M Cain, Camilla Antonucci, Antonio Soliani, John F Clarkin, and Fabio Sambataro. Facial emotion recognition and borderline personality pathology. *Psychiatry research*, 255:347–354, 2017.

[23] Fawzi Rida, Liz Rincon Ardila, Luis Enrique Coronado, Amine Nait-ali, and Gentiane Venture. From motion to emotion prediction: A hidden biometrics approach. In *Hidden Biometrics*, pages 185–202. Springer, 2020.

[24] Joochan Kim, Jungryul Seo, and Teemu H Laine. Detecting boredom from eye gaze and eeg. *Biomedical Signal Processing and Control*, 46:302–313, 2018.

[25] J Jessy Jane Christy, S Iwin Thanakumar Joseph, S Sudhakar, and VM Arul Xavier. Emotion detection in text and acoustic based applications. *Journal of Green Engineering*, 10:3006–3020.

[26] John F Rauthmann, David Gallardo-Pujol, Esther M Guillaume, Elysia Todd, Christopher S Nave, Ryne A Sherman, Matthias Ziegler, Ashley Bell Jones, and David C Funder. The situational eight diamonds: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology*, 107(4):677, 2014.

[27] Gillian M Sandstrom, Neal Lathia, Cecilia Mascolo, and Peter J Rentfrow. Putting mood in context: Using smartphones to examine how people feel in different locations. *Journal of Research in Personality*, 69:96–101, 2017.

[28] Jill W Rettberg. *Seeing ourselves through technology: How we use selfies, blogs and wearable devices to see and shape ourselves.* Springer, 2014.

[29] Diane Ketelle. Talking to myself: Diary as a record of life process. *International Journal of Humanities and Social Science*, 2:34–40, 2012.

[30] Marta Roczniewska, Ewelina Smoktunowicz, and Ewa Gruszczyńska. Capturing life and its fluctuations: Experience sampling and daily diary studies in studying within-person variability. *Social Psychological Bulletin*, 15(2):1–7, 2020.

[31] Andrea B Temkin, Jennifer Schild, Avital Falk, and Shannon M Bennett. Mobile apps for youth anxiety disorders: A review of the evidence and forecast of future innovations. *Professional Psychology: Research and Practice*, 51(4):400, 2020.

[32] Cynthia Suveg, Mary Payne, Kristel Thomassin, and Marni L Jacob. Electronic diaries: A feasible method of assessing emotional experiences in youth? *Journal of Psychopathology and Behavioral Assessment*, 32(1):57–67, 2010.

[33] Kristel Thomassin, Diana Morelen, and Cynthia Suveg. Emotion reporting using electronic diaries reduces anxiety symptoms in girls with emotion dysregulation. *Journal of Contemporary Psychotherapy*, 42(4):207–213, 2012.

[34] Mark Connelly, Maggie H Bromberg, Kelly K Anthony, Karen M Gil, Lindsey Franks, and Laura E Schanberg. Emotion regulation predicts pain and functioning in children with juvenile idiopathic arthritis: an electronic diary study. *Journal of pediatric psychology*, 37(1):43–52, 2011.

[35] Emmanouil Chaniotakis, Constantinos Antoniou, and Evangelos Mitsakis. Data for leisure travel demand from social networking services. In *4th hEART Symposium (European Association for Research in Transportation)*, 2015.

[36] Amir Karami, Morgan Lundy, Frank Webb, and Yogesh K Dwivedi. Twitter and research: a systematic literature review through text mining. *IEEE Access*, 8: 67698–67717, 2020.

[37] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*, 2017.

[38] Said A Salloum, Mostafa Al-Emran, Azza Abdel Monem, and Khaled Shaalan. A survey of text mining in social media: facebook and twitter perspectives. *Adv. Sci. Technol. Eng. Syst. J*, 2(1):127–133, 2017.

[39] M Uma Maheswari and Dr JGR Sathiaseelan. Text mining: Survey on techniques and applications. *Int. J. Sci. Res.*, 6(6):45–56, 2017.

[40] Rafael A Calvo, Sidney D'Mello, Jonathan Matthew Gratch, and Arvid Kappas. *The Oxford handbook of affective computing*. Oxford University Press, USA, 2015.

[41] Steve Portman. Reflective journaling: A portal into the virtues of daily writing. *The Reading Teacher*, 73(5):597–602, 2020.

[42] Masumi Iida, Patrick E Shrout, Jean-Philippe Laurenceau, and Niall Bolger. Using diary methods in psychological research. 2012.

[43] Jonathan Gershuny, Teresa Harms, Aiden Doherty, Emma Thomas, Karen Milton, Paul Kelly, and Charlie Foster. Testing self-report time-use diaries against objective instruments in real time. *Sociological Methodology*, 50(1):318–349, 2020.

[44] Austen R Anderson and Blaine J Fowers. Lifestyle behaviors, psychological distress, and well-being: A daily diary study. *Social Science & Medicine*, 263:113263, 2020.

[45] Soomi Lee and David M Almeida. Daily diary design. *The Encyclopedia of Adulthood and Aging*, pages 1–5, 2015.

[46] Lisa Rhee, Joseph B Bayer, and Alex Hedstrom. Experience sampling method. *The International Encyclopedia of Media Psychology*, pages 1–5, 2020.

[47] Rúben Gouveia and Evangelos Karapanos. Footprint tracker: supporting diary studies with lifelogging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2921–2930. ACM, 2013.

[48] Simone Dohle and Wilhelm Hofmann. Consistency and balancing in everyday health behaviour: An ecological momentary assessment approach. *Applied Psychology: Health and Well-Being*, 11(1):148–169, 2019.

[49] Elizabeth Grubgeld. New media, new lives: Self-publication, blogging, performance art. In *Disability and Life Writing in Post-Independence Ireland*, pages 139–166. Springer, 2020.

[50] Sebastian Hammerl, Thomas Hermann, and Helge Ritter. Towards a semi-automatic personal digital diary: detecting daily activities from smartphone sensors. In *Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments*, page 24. ACM, 2012.

[51] Anne Kaun. Open-ended online diaries: Capturing life as it is narrated. *International Journal of Qualitative Methods*, 9(2):133–148, 2010.

[52] Yasuyuki Kawaura, Asako Miura, Kiyomi Yamashita, and Yoshiro Kawakami. From online diary to weblog: Self-expression on the internet. 2010.

[53] Rozalina Bozhilova. Blogs and their social influence. *Journal of Danubian Studies and Research*, 10(1), 2020.

[54] Jonathan Gershuny. Time-use surveys and the measurement of national well-being. *Centre for Time Use Research, University of Oxford, Swansea, UK, Office for National Statistics*, 2011.

[55] Jeremy E Block and Eric D Ragan. Micro-entries: Encouraging deeper evaluation of mental models over time for interactive data systems. *arXiv preprint arXiv:2009.01282*, 2020.

[56] Julie Ménard, Annie Foucreault, Célestine Stevens, Sarah-Geneviève Trépanier, and Paul Flaxman. Daily fluctuations in office-based workers' leisure activities and well-being. *International Journal of Psychological Studies*, 9(1):47, 2016.

[57] Mojtaba Maghrebi, Alireza Abbasi, Taha Hossein Rashidi, and S Travis Waller. Complementing travel diary surveys with twitter data: application of text mining techniques on activity location, type and time. In *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*, pages 208–213. IEEE, 2015.

[58] Alireza Abbasi, Taha Hossein Rashidi, Mojtaba Maghrebi, and S Travis Waller. Utilising location based social media in travel survey methods: bringing twitter

data into the play. In *Proceedings of the 8th ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, page 1. ACM, 2015.

[59] Niket Tandon, Gerard De Melo, Abir De, and Gerhard Weikum. Knowlywood: Mining activity knowledge from hollywood narratives. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 223–232. ACM, 2015.

[60] Charles T Taylor, Sarah L Pearlstein, Sanskruti Kakaria, Sonja Lyubomirsky, and Murray B Stein. Enhancing social connectedness in anxiety and depression through amplification of positivity: Preliminary treatment outcomes and process of change. *Cognitive Therapy and Research*, pages 1–13, 2020.

[61] Feng Wang, Heather M Orpana, Howard Morrison, Margaret De Groh, Sulan Dai, and Wei Luo. Long-term association between leisure-time physical activity and changes in happiness: analysis of the prospective national population health survey. *American journal of epidemiology*, 176(12):1095–1100, 2012.

[62] Maja Tadic, Wido GM Oerlemans, Arnold B Bakker, and Ruut Veenhoven. Daily activities and happiness in later life: the role of work status. *Journal of happiness studies*, 14(5):1507–1527, 2013.

[63] Tahnee Nicholson and Barbara Griffin. Thank goodness it's friday: weekly pattern of workplace incivility. *Anxiety, Stress, & Coping*, 30(1):1–14, 2017.

[64] Mahnaz Roshanaei, Richard Han, and Shivakant Mishra. Having fun?: Personalized activity-based mood prediction in social media. In *Prediction and Inference from Social Networks and Social Media*, pages 1–18. Springer, 2017.

[65] Vincent Bremer, Dennis Becker, Burkhardt Funk, and Dirk Lehr. Predicting the individual mood level based on diary data. 2017.

[66] Robert LiKamWa, Yunxin Liu, Nicholas D Lane, and Lin Zhong. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, pages 389–402. ACM, 2013.

[67] Pouneh Soleimaninejadian, Min Zhang, Yiqun Liu, and Shaoping Ma. Mood detection and prediction based on user daily activities. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pages 1–6. IEEE, 2018.

[68] Suwen Lin, Louis Faust, Pablo Robles-Granda, Tomasz Kajdanowicz, and Nitesh V Chawla. Social network structure is predictive of health and wellness. *PloS one*, 14(6):e0217264, 2019.

[69] Liana DesHarnais Bruce, Joshua S Wu, Stuart L Lustig, Daniel W Russell, and Douglas A Nemecek. Loneliness in the united states: A 2018 national panel survey of demographic, structural, cognitive, and behavioral characteristics. *American Journal of Health Promotion*, page 0890117119856551, 2019.

[70] Ed Diener and Martin EP Seligman. Very happy people. *Psychological science*, 13(1):81–84, 2002.

[71] Shiyu Zhang, Laura Baams, Daphne van de Bongardt, and Judith Semon Dubas. Intra-and inter-individual differences in adolescent depressive mood: The role of relationships with parents and friends. *Journal of abnormal child psychology*, 46 (4):811–824, 2018.

[72] Kalevi M Korpela, Terry Hartig, Florian G Kaiser, and Urs Fuhrer. Restorative experience and self-regulation in favorite places. *Environment and behavior*, 33 (4):572–589, 2001.

[73] Kalevi M Korpela and Matti Ylen. Perceived health is associated with visiting natural favourite places in the vicinity. *Health & Place*, 13(1):138–151, 2007.

[74] Paul Ekman, Wallace V Friesen, Maureen O'sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712, 1987.

[75] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4): 169–200, 1992.

[76] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[77] Robert Plutchik. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier, 1980.

[78] Silvan S Tomkins. Affect theory. *Approaches to emotion*, 163(163-195), 1984.

[79] Carroll E Izard. *Human emotions*. Springer Science & Business Media, 2013.

[80] Maryam Hasan, Elke Rundensteiner, and Emmanuel Agu. Automatic emotion detection in text streams by analyzing twitter data. *International Journal of Data Science and Analytics*, 7(1):35–51, 2019.

[81] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, volume 6, pages 417–422. Citeseer, 2006.

[82] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[83] David Watson and Auke Tellegen. Toward a consensual structure of mood. *Psychological bulletin*, 98(2):219, 1985.

[84] Klaus R Scherer. Emotion as a multicomponent process: A model and some cross-cultural data. *Review of personality & social psychology*, 1984.

[85] James A Russell and Lisa Feldman Barrett. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, 76(5):805, 1999.

[86] W Gerrod Parrott. *Emotions in social psychology: Essential readings*. Psychology Press, 2001.

[87] Carlo Strapparava, Alessandro Valitutti, et al. Wordnet affect: an affective extension of wordnet. In *Lrec*, volume 4, page 40. Citeseer, 2004.

[88] Cataldo Musto, Giovanni Semeraro, and Marco Polignano. A comparison of lexicon-based approaches for sentiment analysis of microblog posts. In *DART@ AI* IA*, pages 59–68, 2014.

[89] Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.

[90] Priyanka C Nair, Deepa Gupta, and Bhagavatula Indira Devi. A survey of text mining approaches, techniques, and tools on discharge summaries. In *Advances in Computational Intelligence and Communication Technology*, pages 331–348. Springer, 2020.

[91] Rizwana Irfan, Christine K King, Daniel Grages, Sam Ewen, Samee U Khan, Sajjad A Madani, Joanna Kolodziej, Lizhe Wang, Dan Chen, Ammar Rayes, et al. A survey on text mining in social networks. *The Knowledge Engineering Review*, 30(2):157–170, 2015.

[92] Alexandra Amado, Paulo Cortez, Paulo Rita, and Sérgio Moro. Research trends on big data in marketing: A text mining and topic modeling based literature analysis. *European Research on Management and Business Economics*, 24(1): 1–7, 2018.

[93] Ram Gopal, James R Marsden, and Jan Vanthienen. Information mining—reflections on recent advancements and the road ahead in data, text, and media mining, 2011.

[94] Kirill Sarachuk and Missler-Behr Magdalena. Ict, economic effects and business patterns: A text-mining of existing literature. In *Proceedings of the 2020 the 3rd International Conference on Computers in Management and Business*, pages 40–45, 2020.

[95] Andres Azqueta Gavaldon. *Text-mining in macroeconomics: the wealth of words.* PhD thesis, University of Glasgow, 2020.

[96] Sylvia A van Laar, Kim B Gombert-Handoko, Henk-Jan Guchelaar, and Juliëtte Zwaveling. An electronic health record text mining tool to collect real-world drug treatment outcomes: A validation study in patients with metastatic renal cell carcinoma. *Clinical Pharmacology & Therapeutics*, 108(3):644–652, 2020.

[97] Ritu Agarwal and Vasant Dhar. Big data, data science, and analytics: The opportunity and challenge for is research, 2014.

[98] Anne Humeau-Heurtier. Texture feature extraction methods: A survey. *IEEE Access*, 7:8975–9000, 2019.

[99] Jose L Martinez-Rodriguez, Aidan Hogan, and Ivan Lopez-Arevalo. Information extraction meets the semantic web: a survey. *Semantic Web*, (Preprint):1–81, 2020.

[100] Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, and Juan Miguel Gómez-Berbís. Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5):482–489, 2013.

[101] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[102] Yan Wen, Cong Fan, Geng Chen, Xin Chen, and Ming Chen. A survey on named entity recognition. In *International Conference in Communications, Signal Processing, and Systems*, pages 1803–1810. Springer, 2019.

[103] Archana Goyal, Vishal Gupta, and Manish Kumar. Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, 29:21–43, 2018.

[104] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244, 1990.

[105] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[106] Christopher SG Khoo and Sathik Basha Johnkhan. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4):491–511, 2018.

[107] Itisha Gupta and Nisheeth Joshi. Real-time twitter corpus labelling using automatic clustering approach. *International Journal of Computing and Digital Systems*, 9:1–9, 2020.

[108] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[109] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[110] Raghavendra Vijay Bhasker Vangara, Shiva Prasad Vangara, and VR Kailashnath Thirupathur. A survey on natural language processing in context with machine learning, 2020.

[111] Chenglong Ma, Weiqun Xu, Peijia Li, and Yonghong Yan. Distributional representations of words for short text classification. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 33–38, 2015.

[112] Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 165–174. ACM, 2016.

[113] Pratap Chandra Sen, Mahimarnab Hajra, and Mitadru Ghosh. Supervised classification algorithms in machine learning: A survey and review. In *Emerging Technology in Modelling and Graphics*, pages 99–111. Springer, 2020.

[114] Sattam Almatarneh and Pablo Gamallo. Comparing supervised machine learning strategies and linguistic features to search for very negative opinions. *Information*, 10(1):16, 2019.

[115] Atif Khan, Muhammad Adnan Gul, M Irfan Uddin, Syed Atif Ali Shah, Shafiq Ahmad, Al Firdausi, Muhammad Dzulqarnain, and Mazen Zaindin. Summarizing online movie reviews: a machine learning approach to big data analytics. *Scientific Programming*, 2020, 2020.

[116] Ammara Zamir, Hikmat Ullah Khan, Waqar Mehmood, Tassawar Iqbal, and Abubakker Usman Akram. A feature-centric spam email detection model using diverse supervised machine learning algorithms. *The Electronic Library*, 2020.

[117] Inoshika Dilrukshi, Kasun De Zoysa, and Amitha Caldera. Twitter news classification using svm. In *2013 8th International Conference on Computer Science & Education*, pages 287–291. IEEE, 2013.

[118] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 841–842. ACM, 2010.

[119] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary. Twitter trending topic classification. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 251–258. IEEE, 2011.

[120] Yoko Nishihara, Keita Sato, and Wataru Sunayama. Event extraction and visualization for obtaining personal experiences from blogs. In *Symposium on Human Interface*, pages 315–324. Springer, 2009.

[121] Chunyong Yin, Jun Xiang, Hui Zhang, Jin Wang, Zhichao Yin, and Jeong-Uk Kim. A new svm method for short text classification based on semi-supervised learning. In *Advanced Information Technology and Sensor Application (AITS), 2015 4th International Conference on*, pages 100–103. IEEE, 2015.

[122] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[123] Ammar Ismael Kadhim. Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, 52(1):273–292, 2019.

[124] Amanpreet Singh, Narina Thakur, and Aakanksha Sharma. A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1310–1315. Ieee, 2016.

[125] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[126] Fasheng Liu and Lu Xiong. Survey on text clustering algorithm. In *2011 IEEE 2nd International Conference on Software Engineering and Service Science*, pages 901–904. IEEE, 2011.

[127] Neha Garg and RK Gupta. Clustering techniques on text mining: A review. *International Journal of Engineering Research*, 5(4):241–243, 2016.

[128] A Kousar Nikhath and K Subrahmanyam. Feature selection, optimization and clustering strategies of text documents. *International Journal of Electrical & Computer Engineering (2088-8708)*, 9(2), 2019.

[129] Xiangfeng Dai, Marwan Bikdash, and Bradley Meyer. From social media to public health surveillance: Word embedding based clustering method for twitter classification. In *SoutheastCon 2017*, pages 1–7. IEEE, 2017.

[130] Han Kyul Kim, Hyunjoong Kim, and Sungzoon Cho. Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing*, 266:336–352, 2017.

[131] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

[132] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[133] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, 2016.

[134] Diane Cook, Kyle D Feuz, and Narayanan C Krishnan. Transfer learning for activity recognition: A survey. *Knowledge and information systems*, 36(3):537–556, 2013.

[135] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[136] Orly Moreno, Bracha Shapira, Lior Rokach, and Guy Shani. Talmud: transfer learning for multiple domains. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 425–434. ACM, 2012.

[137] Ou Jin, Nathan N Liu, Kai Zhao, Yong Yu, and Qiang Yang. Transferring topical knowledge from auxiliary long texts for short text clustering. In *Proceedings of the*

*20th ACM international conference on Information and knowledge management*, pages 775–784. ACM, 2011.

[138] Aynur Dayanik, David D Lewis, David Madigan, Vladimir Menkov, and Alexander Genkin. Constructing informative prior distributions from domain knowledge in text classification. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 493–500. ACM, 2006.

[139] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447, 2007.

[140] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Boosting for transfer learning. In *Proceedings of the 24th international conference on Machine learning*, pages 193–200. ACM, 2007.

[141] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Self-taught clustering. In *Proceedings of the 25th international conference on Machine learning*, pages 200–207. ACM, 2008.

[142] Chang Wang and Sridhar Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[143] Mingsheng Long, Jianmin Wang, Guiguang Ding, Sinno Jialin Pan, and S Yu Philip. Adaptation regularization: A general framework for transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 26(5):1076–1089, 2013.

[144] Taranpreet Singh Saini, Mangesh Bedekar, and Saniya Zahoor. Analysing human feelings by affective computing-survey. In *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*, pages 1–6. IEEE, 2016.

[145] Armin Seyeditabari, Narges Tabari, and Wlodek Zadrozny. Emotion detection in text: a review. *arXiv preprint arXiv:1806.00674*, 2018.

[146] Edward Chao-Chun Kao, Chun-Chieh Liu, Ting-Hao Yang, Chang-Tai Hsieh, and Von-Wun Soo. Towards text-based emotion detection a survey and possible improvements. In *2009 International Conference on Information Management and Engineering*, pages 70–74. IEEE, 2009.

[147] Ruchi Hirat and Namita Mittal. A survey on emotion detection techniques using text in blogposts. *International Bulletin of Mathematical Research*, 2(1):180–187, 2015.

[148] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 579–586. Association for Computational Linguistics, 2005.

[149] Jerome R Bellegarda. Emotion analysis using latent affective folding and embedding. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 1–9. Association for Computational Linguistics, 2010.

[150] Maryam Hasan, Elke Rundensteiner, and Emmanuel Agu. Emotex: Detecting emotions in twitter messages. 2014.

[151] Alistair Kennedy and Diana Inkpen. Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, 22(2):110–125, 2006.

[152] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[153] Yajie Hu, Xiaoou Chen, and Deshun Yang. Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *ISMIR*, pages 123–128, 2009.

[154] Shuai Yuan, Huan Huang, and Linjing Wu. Use of word clustering to improve emotion recognition from short text. *Journal of Computing Science and Engineering*, 10(4):103–110, 2016.

[155] Shi Feng, Daling Wang, Ge Yu, Wei Gao, and Kam-Fai Wong. Extracting common emotions from blogs based on fine-grained sentiment clustering. *Knowledge and information systems*, 27(2):281–302, 2011.

[156] Mary Jane C Samonte, Hector Irvin B Punzalan, Richard Julian Paul G Santiago, and Peter Joshua L Linchangco. Emotion detection in blog posts using keyword spotting and semantic analysis. In *Proceedings of the 3rd International Conference on Communication and Information Processing*, pages 6–13, 2017.

[157] Shiv Naresh Shivhare, Shakun Garg, and Anitesh Mishra. Emotionfinder: Detecting emotion from blogs and textual documents. In *International Conference on Computing, Communication & Automation*, pages 52–57. IEEE, 2015.

[158] Michael Chau, Tim MH Li, Paul WC Wong, Jennifer J Xu, Paul SF Yip, and Hsinchun Chen. Finding people with emotional distress in online social media: A design combining machine learning and rule-based classification. *MIS Quarterly*, 44(2), 2020.

[159] Elizabeth White, Liam Basford, Stephen Birch, Alison Black, Alastair Culham, Hazel McGoff, Karsten Lundqvist, Philippa Oppenheimer, Jonathan Tanner, Mark Wells, et al. Creating and implementing a biodiversity recording app for teaching and research in environmental studies. *The Journal of Educational Innovation, Partnership and Change*, 1(1), 2015.

[160] Florian Heigl and Johann G Zaller. Using a citizen science approach in higher education: A case study reporting roadkills in austria. *Hum Comput*, 1(2):163–73, 2014.

[161] Jie Lu, Vahid Behbood, Peng Hao, Hua Zuo, Shan Xue, and Guangquan Zhang. Transfer learning using computational intelligence: a survey. *Knowledge-Based Systems*, 80:14–23, 2015.

[162] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[163] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, 2017.

[164] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[165] Charu C Aggarwal and ChengXiang Zhai. A survey of text clustering algorithms. In *Mining text data*, pages 77–128. Springer, 2012.

[166] Harsha S Gowda, Mahamad Suhil, DS Guru, and Lavanya Narayana Raju. Semi-supervised text categorization using recursive k-means clustering. In *International Conference on Recent Trends in Image Processing and Pattern Recognition*, pages 217–227. Springer, 2016.

[167] Wang Yu and Xu Linying. Research on text categorization of knn based on k-means for class imbalanced problem. In *2016 Sixth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC)*, pages 579–583. IEEE, 2016.

[168] Daniel Godfrey, Caley Johns, Carl Meyer, Shaina Race, and Carol Sadek. A case study in text mining: Interpreting twitter data from world cup tweets. *arXiv preprint arXiv:1408.5427*, 2014.

[169] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press, 2008.

[170] Sariel Har-Peled and Bardia Sadri. How fast is the k-means method? *Algorithmica*, 41(3):185–202, 2005.

[171] Pratap Dangeti. *Statistics for Machine Learning.* Packt Publishing Ltd, 2017.

[172] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[173] Chris Fraley and Adrian E Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The computer journal*, 41(8):578–588, 1998.

[174] Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *LREC*, pages 859–866, 2014.

[175] David Jurgens. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 556–562, 2013.

[176] Daniele Vitale, Paolo Ferragina, and Ugo Scaiella. Classification of short texts by deploying topical annotations. In *European Conference on Information Retrieval*, pages 376–387. Springer, 2012.

[177] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern recognition*, 37(9):1757–1771, 2004.

[178] Paul André, Michael Bernstein, and Kurt Luther. Who gives a tweet?: evaluating microblog content value. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 471–474. ACM, 2012.

[179] Axel Schulz, Christian Guckelsberger, and Frederik Janssen. Semantic abstraction for generalization of tweet classification: An evaluation of incident-related tweets. *Semantic Web*, 8(3):353–372, 2017.

[180] Jesse Freitas and Heng Ji. Identifying news from tweets. In *Proceedings of the first workshop on NLP and computational social science*, pages 11–16, 2016.

[181] Daniel Bär, Torsten Zesch, and Iryna Gurevych. A reflective view on text similarity. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 515–520, 2011.

[182] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

[183] Flavia Cristina Bernardini, Rodrigo Barbosa da Silva, Rodrigo Magalhaes Rodovalho, and Edwin Benito Mitacc Meza. Cardinality and density measures and their influence to multi-label learning methods. *Dens*, 1:1, 2014.

[184] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text classification algorithms: A survey. *Information*, 10(4):150, 2019.

[185] Huibing Wang, Jinbo Xiong, Zhiqiang Yao, Mingwei Lin, and Jun Ren. Research survey on support vector machine. *People's Repub. China*, pages 95–103, 2017.

[186] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2016.

[187] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[188] M Taylor Dryman and Richard G Heimberg. Emotion regulation in social anxiety and depression: A systematic review of expressive suppression and cognitive reappraisal. *Clinical Psychology Review*, 65:17–42, 2018.

[189] Katie Finning, Tamsin Ford, Darren A Moore, and Obioha C Ukoumunne. Emotional disorder and absence from school: findings from the 2004 british child and adolescent mental health survey. *European child & adolescent psychiatry*, 29(2): 187–198, 2020.

[190] Lowri Williams, Michael Arribas-Ayllon, Andreas Artemiou, and Irena Spasić. Comparing the utility of different classification schemes for emotive language analysis. *Journal of Classification*, pages 1–30, 2019.

[191] Laurence Devillers, Laurence Vidrascu, and Lori Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18 (4):407–422, 2005.

[192] Agnieszka Landowska. Towards new mappings between emotion representation models. *Applied Sciences*, 8(2):274, 2018.

[193] Liam D Turner. *Decomposing responses to mobile notifications.* PhD thesis, Cardiff University, 2017.

[194] Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

[195] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[196] Mahendra Sahare and Hitesh Gupta. A review of multi-class classification for imbalanced data. *International Journal of Advanced Computer Research*, 2(3): 160, 2012.

[197] Christie Napa Scollon, Chu-Kim Prieto, and Ed Diener. Experience sampling: promises and pitfalls, strength and weaknesses. In *Assessing well-being*, pages 157–180. Springer, 2009.

[198] Andrea Caputo. Social desirability bias in self-reported well-being measures: Evidence from an online survey. *Universitas Psychologica*, 16(2):245–255, 2017.

[199] Nuria Daviu, Michael R Bruchas, Bita Moghaddam, Carmen Sandi, and Anna Beyeler. Neurobiological links between stress and anxiety. *Neurobiology of stress*, 11:100191, 2019.

[200] Jeremy DeMartini, Gayatri Patel, and Tonya L Fancher. Generalized anxiety disorder. *Annals of internal medicine*, 170(7):ITC49–ITC64, 2019.

[201] Jonathan Iwry. Toward a psychological theory of meaning in life. *Available at SSRN 3497017*, 2019.

[202] Martin EP Seligman. Positive psychology: A personal history. *Annual review of clinical psychology*, 15:1–23, 2019.

[203] Paul Ekman. Facial expressions of emotion: New findings, new questions, 1992.

[204] Liam D Turner, Stuart M Allen, and Roger M Whitaker. Interruptibility prediction for ubiquitous systems: conventions and new directions from a growing field. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 801–812, 2015.

[205] Martin J Chorley, Roger M Whitaker, and Stuart M Allen. Personality and location-based social networks. *Computers in Human Behavior*, 46:45–56, 2015.

[206] Ryne A Sherman, Christopher S Nave, and David C Funder. Situational construal is related to personality and gender. *Journal of Research in Personality*, 47(1): 1–14, 2013.