Deep Learning Approach to Sentiment Analysis in Health and Well-Being

A thesis submitted in partial fulfilment

of the requirement for the degree of Doctor of Philosophy

Anastazia Žunić

September 2022

Cardiff University
School of Computer Science & Informatics

To my mother and my grandfather,
for their unconditional support.

Abstract

Sentiment analysis, also known as opinion mining, is an area of natural language processing which focuses on the classification of the sentiment that is expressed in a written document. Sentiment analysis has found applications in various domains including finance, politics, and health. This thesis is focused on sentiment analysis in the domain of health and well-being. An extensive systematic literature review was carried out to establish the state of the art in sentiment analysis in this domain. This systematic review provides evidence that the state-of-the-art results in sentiment analysis in the domain of health and well-being lags behind that in other domains. Additionally, it revealed that deep learning has not been used to classify the sentiment within the aforementioned domain. Furthermore, we performed a study and showed that the language that is used within the health and well-being domain is biased towards the negative sentiment. Aspect-based sentiment analysis refines the focus of sentiment analysis by classifying the sentiment associated with a specific aspect. Subsequently, we focus specifically on aspect-based sentiment analysis. To support it within the domain of health and well-being we created a dataset consisting of drug reviews, where the aspects were automatically annotated by matching concepts from the Unified Medical Language System. We have successfully shown that graph convolution can effectively utilise the context, represented with syntactic dependencies, to determine the intended sentiment of inherently negative aspects and consequently close the performance gap regardless of the domain. The advent of transformer-based architectures initiated a breakthrough in various tasks in natural language processing, including sentiment analysis. Therefore, we presented an approach to fine-tuning a transformer-based language model for the specific task of aspect-based sentiment analysis. The findings show the evidence that transformer-based models account for syntactic dependencies when classifying the sentiment of the given aspect.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor Prof. Irena Spasić for her uncountable support, guidance and the tremendous amount of help throughout these years. For sharing her expertise with me and for teaching me professional and research skills. My utmost gratitude is extended to my second supervisor Dr. Padraig Corcoran for his continuous support, guidance and for sharing his expertise with me throughout my PhD.

I am immensely grateful to Cardiff University for offering me the Vice-Chancellor's International Scholarship for Research Excellence and to Cardiff School of Computer Science and Informatics who have granted me the scholarship for this research project.

To all my friends who I met during this journey and for the wonderful moments we shared, thank you for your constant support. To my friends back home, thank you for always believing in me.

Finally, I would like to thank my family. Especially my mother Silvia who has been a true inspiration in every aspect of my life, I am eternally grateful for your patience, support and love. My greatest thank you goes to my grandfather Coriolan without whom this all could not be possible.

Contents

List of Publications

The work described in this thesis is based on the following publications.

- Anastazia Žunić, Padraig Corcoran, and Irena Spasić. Sentiment analysis in health and well-being: systematic review. *JMIR Medical Informatics*, 8(1), e16023, 2020.

- Anastazia Žunić, Padraig Corcoran, and Irena Spasić. Improving the performance of sentiment analysis in health and wellbeing using domain knowledge. *HealTAC 2020: Healthcare Text Analytics Conference*, London, UK, Online, 2020.

- Anastazia Žunić, Padraig Corcoran, and Irena Spasić. Aspect-based sentiment analysis with graph convolution over syntactic dependencies. *Artificial Intelligence in Medicine*, 119, 102138, 2021.

- Anastazia Žunić, Padraig Corcoran, and Irena Spasić. The case of aspect in sentiment analysis: Seeking attention or co-dependency? *Machine Learning and Knowledge Extraction* 4(2), 474-487, 2022.

List of Figures

List of Tables

List of Acronyms

SA Sentiment Analysis

NLP Natural Language Processing

GCN Graph Convolutional Network

CNN Convolutional Neural Network

LSTM Long Short-Term Memory

BiLSTM Bidirectional Long Short-Term Memory

GloVe Global Vectors for Word Representation

ML Machine Learning

TP True Positive

TN True Negative

FP False Positive

FN False Negative

RNN Recurrent Neural Network

BERT Bidirectional Encoder Representations from Transformers

ELMo Embeddings from Language Models

GRU Gated Recurrent Unit

BiGRU Bidirectional Gated Recurrent Unit

GPT Generative Pre-trained Transformer

SVM Support Vector Machine

CRF Conditional Random Fields

KNN k-Nearest Neighbors

DT Decision Tree

NB Naïve Bayes

ME Maximum Entropy

LR Logistic Regression

RF Random Forest

Chapter 1

Introduction

The rapid growth of the online platforms, where users generate opinionated and informal content, often results in information overload, which calls for automated text-mining solutions to quickly determine the key aspects of public opinion. Natural language processing (NLP) has found its practical applications in efficiently dealing with the ever increasing amount of textual data written in natural languages such as English. Specifically, sentiment analysis (SA), also known as opinion mining, is a subfield of NLP, which focuses on the problem of automatically classifying the public sentiment expressed in free text typically originating from the Internet [161]. Formally, SA is defined as the task of identifying a quadruple (s, g, h, t) whose values represent the sentiment, the object targeted by the sentiment, the holder of the sentiment and the time at which the sentiment was expressed [162]. In practice, SA has traditionally focused on a simpler task of finding the pair (s, g). Here, the sentiment represents a reflection of one's attitudes, feelings, opinions towards specific entities, and it is most often conflated to polarity [63], which can be either positive or negative, even though other classification schemes can be utilised [256]. The target g has typically been the overall topic of an analysed text document. In principle, the target is some entity, but can also be an aspect of such an entity. Here, an aspect represents some characteristic of such an entity [226]. The choice of these characteristics depends on a specific domain in which SA is applied. Aspect-based SA refines the focus of SA by classifying the sentiment associated with a specific aspect and not just the overall sentiment associated with the entity.

The origins of SA can be traced back to the 1990s including methods for classifying the point of view [253], predicting the semantic orientation of adjectives [120], subjectivity classification [254], etc. However, the increased research activities in relation to SA is correlated with the advent of Web 2.0 and, in particular, the proliferation of social media communication channels that has led to the ever so increasing availability of user-generated data. SA has been applied across a wide range of societal contexts including marketing, economy and politics [131, 130, 59, 101, 213].

This thesis focuses specifically on SA related to health and well-being. Health is defined as "a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity" [38]. On the other hand, well-being is considered to be a perceived or subjective state, that is, it can vary considerably across individuals with similar circumstances [134]. Given the prevalence of sentiment information, this makes health and well-being an ideal application domain for SA. Nevertheless, when it comes to matters of health, modern society tends to be preoccupied with the negative phenomena such as diseases, injuries, and disabilities [56], which makes SA in this domain challenging. For instance, consider a patient with a chronic condition. Having a good quality of life will not necessarily depend on the absence of associated symptoms, but rather on the extent to which they are managed and controlled. However, the negative connotation of health symptoms tends to skew the SA results toward the negative spectrum. This problem also creates an interesting research opportunity for further advances in SA from their methodological perspective.

1.1 Motivation

SA has been extensively researched during the last decade [152]. The resulting methods have also found their applications in the domain of health and well-being. A limited availability of clinical data resulting from privacy concerns and the need for manual data annotation [231] has geared the SA application in health and well-being towards user-generated data, which are often publicly available. Typical applications focus on determining how people and patients feel towards their treatment, their overall progress, certain medications, and generally aspects that are related to one's health and well-being. Not surprisingly, much of the research into SA in relation to health and well-being focuses on drug reviews. Aspect-based SA of such reviews can in turn be used to support pharmacovigilance by detecting adverse drug reactions [151]. The most obvious aspects in this case would be drug indications and side effects. For instance, consider the following examples in which the word headache represents an aspect:

1. It's the only drug that works for my $\underbrace{\text{headache}}_{\text{positive}}$.
2. A dose of 750 mg twice daily had no effect on my $\underbrace{\text{headache}}_{\text{negative}}$.
3. Caused vomiting and gave me the worst $\underbrace{\text{headache}}_{\text{negative}}$.
4. I find using a half a capsule seems to work fine without giving me a $\underbrace{\text{headache}}_{\text{positive}}$.

These examples illustrate how the sentiment towards the same aspect can vary across different contexts. Unlike their counterparts in other domains, e.g. quality and price in product reviews, aspects such as headache are a priori negative. The key issue that prevents successful re-use of the existing off-the-shelf SA solutions is that the narratives within this domain are often incorrectly classified into negative polarity. Intuitively, this phenomenon of "false negatives" can be explained by the underlying nature of such narratives whose main topics represent one's symptoms or conditions, e.g. headache, nausea, pain. In general, they are focused around diseases, injuries and disabilities. Therefore, to accurately classify the sentiment associated with otherwise negative aspects requires careful consideration of problem representation and the methods operating on them. Specifically, the interaction of an aspect with its context and the effects these interactions have on its sentiment is at the core of this problem.

We framed the problem of aspect-based SA as a machine learning problem, where the main aim is to automatically learn to detect such interactions and infer the resulting sentiment. Deep learning has been singled out as one class of machine learning algorithms that have the capability of learning complex non-linear binary classifications and as such demonstrate the greatest potential for dealing with the problem of aspect-based SA. Deep learning algorithms involve the use of large multi-layer neural network models. The use of deep learning has led to many scientific breakthroughs in many domains including NLP and computer vision [108]. The use of deep learning in NLP has emerged relatively recently following the advent of word embedding techniques such as *word2vec* [179] in 2013. Word embeddings are based on the concept of distributional semantics whereby words are represented by vectors such that words that appear in similar contexts have similar vectors. Prior to the breakthrough of deep learning, solutions to SA within the domain of health and well-being, included rule-based and statistical methods that operated on the bag-of-words representation and relied on manual feature engineering. However, such approaches do not account for many important aspects of natural language such as the context's structure. Consequently, they are likely to underperform in domains such as health and well-being where the polarity of individual words when they are considered out of context tends to be negative.

1.2 Research Aims and Objectives

Features used to support SA include terms, part of speech, syntactic dependencies and negation [202]. Most commonly, opinionated words that carry subjective bias are used in a bag-of-words approach to classify opinions [49]. Opinionated words can be utilised from lexicons such as SentiWordNet [51] or WordNet-Affect [244]. Other features explored in SA include more complex linguistic models based on lexical substitution, n-grams and phrases [84, 228].

The main hypothesis of this project is that syntactic dependencies can improve the performance of aspect-based SA. This hypothesis naturally leads us to the following research question:

How can syntactic dependencies be utilised as features to improve the performance of aspect-based SA?

To answer this question, we identified the following research objectives:

RO1. The first step towards incorporating syntactic dependencies as features of aspect-based SA is to develop a suitable problem representation. Therefore, our first objective is to develop a representation that integrates syntactic dependencies and vector representation of individual words whilst preserving all information contained in the original raw text. Here, we do not necessarily commit to explicit representation of syntactic dependencies and leave room to later explore deep learning approaches that can learn syntactic dependencies from data and use them as latent features.

RO2. Once a suitable problem representation has been found, the next objective is to develop a neural network architecture that can effectively utilise such representation. Given the no free lunch theorem [259], we know that this requires a systematic exploration of available deep learning approaches. We will focus on a subset of deep learning approaches that have demonstrated the greatest potential in text mining including recurrent neural networks, convolutional neural networks and transformers.

RO3. The final objective of this project is to use the findings from RO1 and RO2 to refine the proposed neural network architectures, optimise their hyperparameters and develop a suitable training dataset in a quest to improve the state of the art in aspect-based SA in health and well-being.

1.3 Research Contributions

As described previously, this research is motivated by the strong negative polarity associated with the concepts related to the domain of health and well-being and the potential of deep learning models to overcome this challenge. This negative bias was observed when concepts from the largest biomedical terminology were cross-referenced against a set of popular sentiment lexicons [286]. The ratio between positive and negative biomedical concepts was found to be 6:10 on average. In order to establish the state of the art within SA in health and well-being, a systematic literature review was conducted [286]. This systematic literature review is presented in Chapter 2. The findings of the systematic literature review revealed that SA in health and well-being lags behind the state of the art in other domains such as customer satisfaction with products and services. For example, the accuracy of SA of movie reviews is typically well over 90%, which is close to human-level performance. However, the accuracy of SA in health and well-being typically ranges from 70% to 80%, and sometimes even falls to as low as 60%, which is well below human-level performance. This suggests the need for novel methods for SA, which would perform well in this as well as other domains.

Following the findings from the systematic literature review where the results of SA were obtained on different datasets, we performed a comparative analysis of five publicly available SA tools to gauge their performance on the same dataset related to health and well-being and compare these results to their performance on another dataset, which falls outside of this domain [285]. This study, described in Chapter 3, was inspired by the sublanguage theory of Zellig Harris [117], which purports that a language restricted to a specific domain can be delineated from the language in general in terms of its content and structure. This theory may explain the fact that SA suffers from domain dependency [63]. Within this study we hypothesised the following: (1) the sublanguage of the user-generated content varies across domains, (2) the sublanguage of health and well-being is biased towards the negative sentiment, and (3) medical knowledge can be used to address the bias towards the negative end, and consequently improve the overall performance of SA. Herein we demonstrated that a sublanguage related to health and well-being, specifically the sublanguage of drug reviews, differs from the more general-domain language that is used within the domain of movie reviews. These two sublanguages differ with respect to their vocabulary which was cross-referenced using the Unified Medical Language System (UMLS), a large repository of inter-related biomedical concepts [58]. Following this, the predicted sentiment of the considered off-the-shelf tools was found to be biased towards the negative sentiment. However, following the abstraction of biomedical concepts using the knowledge encoded within the UMLS, the negative bias was reduced and in turn the performance of the SA tools was improved. This finding was in line with those arising from the systematic literature review.

Going back to the systematic literature review, it made it apparent that much of the SA research in the domain of health and well-being failed to utilise deep learning. This was in contrast with the general trends in SA research in other domains, which take advantage of deep learning to deal with the complexity of the given task. This can be attributed to the difficulties of applying deep learning in the domain of health and well-being. Deep learning approaches are data hungry in the sense that their performance is strongly correlated with the amount of training data available and such data may not be abundant in this domain [231]. In addition, pre-trained word embeddings, that are commonly used as input to these models, are typically trained using data containing more general-domain language, such as Google News [179]. Hence, these word embeddings do not capture the true meaning of language used in the domain of health and well-being. However, in this thesis we hypothesise that this challenge may be tackled by considering the syntactic structure of the sentences to contextualise word representation and, specifically, their sentiment.

A sentence can be represented in the form of a graph where vertices correspond to words and edges correspond to their syntactic relationships. In this thesis, we propose a novel neural network architecture based on graph convolution which is applied directly to a dependency parse tree to perform aspect based SA [287]. This work is described in Chapter 5. We demonstrate that this novel approach improves the performance of aspect-based SA in health and well-being relative to other standard neural network architectures that have been previously used in NLP. Specifically, using drug reviews, we created a new dataset that is suitable for aspect-based SA evaluation in the domain of health and well-being. The experimental results on this dataset show that graph convolution achieves the best results. These results in turn suggest that the features that are incorporated within the dependency parse tree of a sentence carry important information for the classification of sentiment.

Recently, NLP has experienced a great breakthrough with the appearance of a new neural network architecture entitled the transformer [245]. The architecture has been used to successfully train very large language models that produce contextualised word embeddings. There exists a large array of neural network architectures based on the original transformer architecture. One of the most popular in the domain of NLP is called Bidirectional Encoder Representations from Transformers (BERT) [92]. The main advantage of BERT is that it not only provides contextualised word embeddings but it can also be easily fine-tuned using relatively small datasets to support downstream NLP tasks such as SA. This is of particular relevance to the domain of health and well-being where the training datasets tend to be on the smaller side [231].

In the final study presented in Chapter 6, we investigate the potential of the BERT architecture for aspect-based SA in health and well-being [288]. We find that this approach outperforms our previous solution based on graph convolution described in Chapter 5. To help understand the reasons for the improved performance, we provide an in-depth analysis of this model and investigate whether it learns to use syntactic dependencies when performing aspect-based SA. We consider an approach to model interpretation based on attribution scores, where these scores represent the relevance of each token within the model. We hypothesise that, if the model uses syntactic dependencies when performing aspect-based SA, there will exist a negative correlation between attribution scores and the distance of tokens to the aspect in the dependency parse tree. The analysis shows that the correlation is indeed negative and this provides evidence that BERT does account for syntactic dependencies when performing aspect-based SA.

In summary, the key research contributions of this thesis are as follows:

- Two novel deep learning approaches, based on graph convolution and transformers respectively, were developed for the problem of aspect-based SA.

- Our experiments confirmed that consideration of syntactic dependencies does improve the performance of aspect-based SA.

- In both cases, we achieved results that have pushed the state of the art in a particularly challenging domain of health and well-being.

1.4 Thesis Structure

The remaining Chapters of the thesis are organised as follows:

- **Chapter 2** - *Background: A Systematic Review of Sentiment Analysis in Health and Well-Being* - Provides an in-depth systematic literature review of sentiment analysis in health and well-being. This review identified the research gap to motivate the research conducted in this PhD project.

- **Chapter 3** - *Measuring the Performance of Sentiment Analysis Tools* - Investigates the performance of off-the-shelf sentiment analysis tools and their performance on health and well-being related content. This case study provides specific evidence that confirms the research gap.

- **Chapter 4** - *An Overview of Deep Learning* - Introduces basic concepts related to deep learning to facilitate the reading of remaining Chapters. It also provides a literature overview of deep learning approaches in sentiment analysis to inform the selection of methods used in our approaches to aspect-based sentiment analysis.

- **Chapter 5** - *A Graph Convolutional Approach to Aspect-Based Sentiment Analysis* - Describes a novel convolutional model for aspect-based sentiment analysis that utilises explicitly encoded syntactic dependencies.

- **Chapter 6** - *A Transformer-Based Approach to Aspect-Based Sentiment Analysis* - Describes a novel model for aspect-based sentiment analysis based on transformers adapted for aspect-based sentiment analysis. We extended the analysis to provide evidence that suggests the model has learnt to take advantage of implicitly encoded syntactic dependencies.

- **Chapter 7** - *Conclusion* - Concludes the thesis and highlights opportunities for future work.

Chapter 2

Background: A Systematic Review of Sentiment Analysis in Health and Well-Being

The work presented in this Chapter has been published in JMIR Medical Informatics. It is based on the following paper: Žunić Anastazia, Corcoran Padraig and Spasić Irena. Sentiment analysis in health and well-being: systematic review. *JMIR Medical Informatics*, 8(1), e16023, 2020.

2.1 Introduction

To be able to establish the state of the art in sentiment analysis (SA) related to health and well-being, the systematic literature review of the recent literature has been conducted. To capture the perspective of those individuals whose health and well-being are affected, the focus of this systematic literature review is specifically on spontaneously generated content and not necessarily that of health care professionals. This differentiates this review from others conducted on related topics. For example, Denecke and Deng [89] reviewed SA in medical settings, but focused on the word usage and sentiment distribution of clinical data, such as nurse letters, radiology reports, and discharge summaries, while public data shared by the likes of patients and caregivers were restricted to 2 websites. On the contrary, Gohil et al. [107] dealt with user-generated data, but only considered Twitter, whereas herein there are no posed restrictions on the platforms used to generate the data.

This Chapter is organised as follows. Section 2.2 explains the methodology of this systematic review in detail. The findings of the review are presented in Section 2.3, this is followed by a discussion, which is given in Section 2.4. The main findings of the systematic literature review are summarised in Section 2.5.

2.2 Methods

This section provides a framework of how the systematic literature review has been carried out.

2.2.1 Guidelines

The methodology of this systematic literature review is based on the guidelines for performing systematic literature reviews described by Kitchenham in [149]. It is structured around the following steps:

1. *Research questions* define the scope, depth, and the overall aim of the review.
2. *Search strategy* is an organised process designed to identify all the studies that are relevant to the research questions in an efficient and reproducible manner.
3. *Inclusion and exclusion criteria* define the scope of a systematic literature review.
4. *Quality assessment* refers to a critical appraisal of included studies to ensure that the findings of the review are valid.
5. *Data extraction* is the process of identifying the relevant information from the included studies.

ID	Question
RQ1	What are the major sources of data?
RQ2	What is the originally intended purpose of spontaneously generated narratives?
RQ3	What are the roles of their authors within health and care?
RQ4	What are their demographic characteristics?
RQ5	What areas of health and well-being are discussed?
RQ6	What are the practical applications of SA?
RQ7	What methods have been used to perform SA?
RQ8	What is the state-of-the-art performance of SA?
RQ9	What resources are available to support SA related to health and well-being?

Table 2.1: Research questions for the systematic literature review.

6. *Data synthesis* involves critical appraisal and synthesis of evidence to support the findings of the review.

2.2.2 Research questions

The overarching topic of this systematic literature review is the SA of spontaneously generated narratives in relation to health and well-being, which have not been created by health care professionals. The main aim is to answer the research questions given in Table 2.1.

2.2.3 Search strategy

To systematically identify articles relevant to SA related to health and well-being, appropriate data sources were considered: the Cochrane Library [6], MEDLINE [20], EMBASE [12], and CINAHL [11]. MEDLINE is chosen as the most diverse data source with respect to the topics covered and publication types. MEDLINE is a premier bibliographic database that contains more than 26 million of references to articles in life sciences and biomedicine. Its coverage dates back to 1946, and its content is updated daily. It covers publications of various types, for example, journal articles, case reports, conference papers, letters, comments, guidelines, and clinical trials. Its content is systematically indexed by Medical Subject Headings (MeSH), a hierarchically organised terminology for cataloguing biomedical information, to facilitate identification of relevant articles. For example, it defines the term *natural language processing* as “computer processing of a language with rules that reflect and describe current usage rather than prescribed usage”. Therefore, this term can be used to identify articles on this topic even when they use alternative terminology, for example, “sentiment analysis,” “information retrieval,” and “text mining”. We used PubMed, a multifaceted interface, to search MEDLINE.

Next step in developing the search strategy is to define a search query that adequately describes the chosen topic, which is SA related to health and well-being. Given the MEDLINE’s focus on biomedicine, inclusion of terms related to health and well-being is considered redundant. Specifically, they could improve the precision of the search (i.e., reduce the number of irrelevant articles retrieved), but could only decrease the recall (the number of relevant articles retrieved). Given the relative recency of research into SA and its applications in biomedicine, it is expected to create a query focusing solely on SA to retrieve a manageable number of articles, which could then be reviewed manually. The search query is defined as follows:

((sentiment[Title] OR sentiments[Title] OR opinion[Title] OR opinions[Title] OR emotion[Title] OR emotions[Title] OR emotive[Title] OR affect[Title] OR affects[Title] OR affective[Title]) AND (“sentiment classification” OR “opinion mining” OR “natural language processing” OR NLP OR “text analytics” OR “text mining” OR “F-measure” OR “emotion classification”)) OR “sentiment analysis”.

Do note that this search query will also retrieve any references to aspect-based SA as this term explicitly references its hypernym - SA, which is included in the search query in its full form - sentiment analysis. The search performed on January 24, 2019,

ID	Criterion
IN1	The input text represents spontaneously generated narrative.
IN2	The input text discusses topics related to health and well-being.
IN3	The input text captures the perspective of an individual personally affected by issues related to health and well-being (e.g., patient or carer) rather than that of a health care professional.
IN4	Sentiment is analysed automatically using natural language processing.

Table 2.2: Inclusion criteria.

retrieved a total of 299 articles. Notably, no articles published before 2011 were retrieved, which confirmed the hypothesis about the relative recency of research into SA and its applications in biomedicine.

Note: The search was performed again in order to include the articles until January 2021. The originally published systematic literature review [286] has been updated to bring the review up to date, this resulted in a total of 23 added articles [106, 80, 234, 186, 113, 133, 211, 118, 165, 172, 105, 277, 91, 168, 177, 247, 183, 69, 206, 41, 128, 224, 185].

2.2.4 Selection criteria

To further refine the scope of this systematic literature review, a set of inclusion and exclusion criteria are defined (see Tables 2.2 and 2.3) to select the most appropriate articles from those matching the search query. Two annotators independently screened the retrieved articles against inclusion and exclusion criteria and achieved the inter-annotator agreement of 0.51 calculated using Cohen kappa coefficient [78]. Disagreements were resolved by the third independent annotator. A total of 95 articles were retained for further processing.

To ensure the rigorousness and credibility of selected studies, they were additionally evaluated against the quality assessment criteria defined in Table 2.4. A total of 9

ID	Criterion
EX1	Sentiment analysis is performed in a language other than English.
EX2	The article is written in a language other than English.
EX3	The article is not peer reviewed.
EX4	The article does not describe an original study.
EX5	The article was published before January 1, 2000.
EX6	The full text of the article is not freely available to the academic community.

Table 2.3: Exclusion criteria.

ID	Criterion
QA1	Are the aims of the research clearly defined?
QA2	Is the study methodologically sound?
QA3	Is the method explained in sufficient detail to reproduce the results?
QA4	Were the results evaluated systematically?

Table 2.4: Quality assessment criteria.

studies were found not to match the given criteria. This further reduced the number of selected articles to 86. Figure 2.1 summarises the outcomes of the 4 major stages in the systematic literature review. As mentioned before, a total of 23 articles were additionally added to bring the systematic literature up to date.

2.2.5 Data extraction and synthesis

Data extraction cards provide support for the collection of information that are relevant to the research questions. They included items described in Table 2.5. The selected articles were read in full to populate the data extraction cards, which were then used to facilitate narrative synthesis of the main findings.

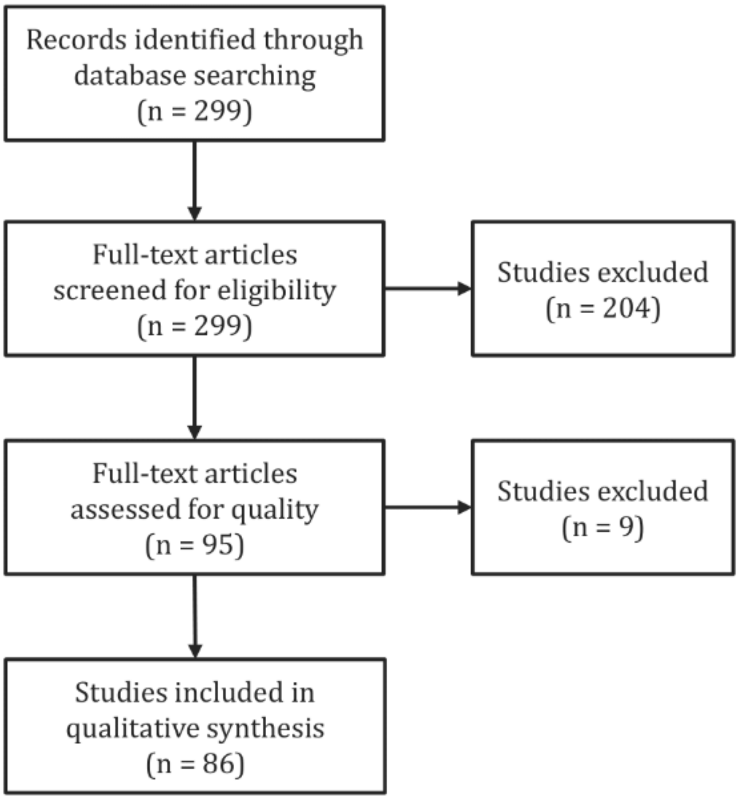
Records identified through database searching (n = 299)
Full-text articles screened for eligibility (n = 299)
Studies excluded (n = 204)
Full-text articles assessed for quality (n = 95)
Studies excluded (n = 9)
Studies included in qualitative synthesis (n = 86)

Figure 2.1: Flow diagram of the systematic literature review process before it has been updated for the purpose of the thesis.

Item	Description
Data	Provenance, purpose, selection criteria, size, and use.
Topic	General topic discussed in the given dataset including medical conditions and treatments.
Author	Author (data creator) demographics and their role in health and care.
Application	Downstream application of SA results.
Method	Type of SA method used, feature selection/extraction, and any resources used to support implementation of the method.
Evaluation	Measures used to evaluate the results, specific results reported, baseline method used, and improvements over the baseline (if any).

Table 2.5: Data extraction framework.

2.3 Results

Findings of this systematic literature review are presented in this section.

2.3.1 Data provenance

This section discusses the answers to RQ1 and to RQ2 which focus on the main properties of data used as input for SA. The majority of data were collected from the mainstream social multimedia and Web-based retailing platforms, which provide the most pervasive user base together with application programming interfaces (APIs) that can support large-scale data collection. Not surprisingly, 39 studies [275, 182, 212, 255, 241, 284, 82, 151, 194, 201, 57, 85, 98, 97, 112, 114, 142, 160, 178, 191, 196, 222, 64, 104, 199, 278, 106, 211, 172, 105, 177, 247, 183, 206, 41, 128, 224, 185, 277] used data sourced from Twitter, a social networking service on which users post messages restricted to 280 characters (previously 140). Twitter can be accessed via its API from a range of popular programming languages using libraries such as TwitterR [212], Twitter4J in Java [201, 222], and Tweepy in Python [278].

Facebook, another social networking service, was used to collect user posts regarding Chron's disease [219] and depression and anxiety [137]. Comments posted on Instagram, a photo and video-sharing social networking service, were used to predict depression [217]. A total of 2 studies used data from YouTube, a video-sharing website, which allows users to share videos and comment on them. These studies collected comments on videos related to pro-anorexia [195] and Invisalign experience [166]. Reddit, a social news aggregation, Web content rating, and discussion website, was used to learn to differentiate between suicidal and non-suicidal comments [43]. Amazon, a Web-based retailer, allows users to submit reviews of products. Customers may comment or vote on the reviews, much in the spirit of social networking websites. Amazon is the largest single source of consumer reviews on the internet. Amazon reviews were collected from the section of joint and muscle pain relief treatments [40].

Mainstream social media provides a generic platform to engage patients. One of their advantages in this context is that many patients are already active users of these platforms, thus effectively lowering the barrier to entry to engaging patients online. However, the use of social media in the context of disclosing protected health information may raise ethical issues such as those related to confidence and privacy. The need to engage patients online while fully complying with data protection regulations has led to the proliferation of websites and networks developed specifically to provide a safe space for sharing health-related information online. This systematic review identified 11 platforms of this kind that have been utilised in 29 studies, details can be seen in Table 2.6.

Due to ethical concerns, the data used in these studies are usually not released publicly to support further research and evaluation. Only one such dataset has been published. The eDiseases dataset used in 2 studies [67, 65] contains patient data from the MedHelp website (see Table 2.6). The dataset contains 10 conversations from 3 patient communities, allergies, Crohn's disease, and breast cancer, which according to a medical expert, exhibit high degree of heterogeneity with respect to health literacy and demographics. The conversations were selected randomly out of those that contained at least 10 user posts. Individual sentences were annotated with respect to their factuality (opinion, fact, or experience) and polarity (positive, negative, or neutral). Annotation was performed by 3 frequent users of health forums. With approximately 3000 annotated sentences with high degree of heterogeneity, this dataset represents a suitable testbed for evaluating SA in the health domain.

As illustrated by the studies discussed thus far, spontaneously generated narrative used in SA typically coincides with the user-generated content, that is, content created by a user of an online platform and made publicly available to other users. The fifth i2b2/VA/Cincinnati challenge in NLP for clinical data [207] represents an important milestone in SA research related to health and well-being. The challenge focused on the task of classifying emotions from suicide notes. The corpus used for this shared

Website	Description	Used in
RateMDs [26]	Allows users to post reviews about health care staff and services.	[45, 246, 126]
WebMD [34]	Publishes content about health and care topics, including fora that allow users to create and participate in support groups and discussions.	[255, 135, 189, 113]
Ask a Patient [3]	Allows users to share their personal experience about drug treatments.	[189, 190, 165]
DrugLib.com [9], Drugs.com [10]	Allows users to rate and review prescription drugs.	[255, 189, 190, 47, 80]
Breastcancer.org [4]	A breast cancer community, where members discuss various topics related to breast cancer.	[279, 62]
MedHelp [17]	Allows users to share their personal experiences and evidence-based information related to health and well-being.	[182, 67, 65, 270, 169]
DailyStrength [7]	A social networking service that allows users to create support groups across different categories related to health and well-being.	[255, 151]
Cancer Survivors Network [5]	A social networking service that connects users whose lives have been affected by cancer and allows them to share personal experience and expressions of caring.	[208, 282, 61]
NHS website (formerly NHS Choices) [19]	The primary public facing website of the United Kingdom's National Health Service (NHS) with more than 43 million visits per month. It provides health-related information and allows patients to provide feedback on services.	[111]
DiabetesDaily [8]	A social networking service that connects people affected by diabetes where they can trade advice and learn more about the condition.	[42]

Table 2.6: Health-related websites and networks.

task contained 1319 written notes left behind by people who died by suicide. Individual sentences were annotated with the following labels: abuse, anger, blame, fear, guilt, hopelessness, sorrow, forgiveness, happiness, peacefulness, hopefulness, love, pride, thankfulness, instructions, and information. A total of 24 teams used these data to develop their classification systems and evaluate their performance, out of which 19 teams published their results [73, 90, 100, 158, 173, 176, 188, 200, 204, 216, 218, 229, 230, 248, 252, 268, 271, 273, 276].

The vast majority of data used in studies encompassed by this review represent user-generated content originating from online platforms. There are 2 main types of user-generated content: customer reviews and user comments. A customer review is a review of a product or service made by someone who purchased, used, or had experience with the product or service. The main class of products reviewed in the datasets considered here are medicinal products. Product reviews were collected from Amazon, but also from specialised websites such as Ask a Patient, DrugLib.com, and Drugs.com. These reviews provide users with additional information about a product's efficacy and possible side effects typically described in layman's terms, thus lowering a barrier to participation in health care linked to health literacy and potentially providing better support for shared decision making. Other websites such as RateMDs and the National Health Service (NHS) website allow users to review health care services they received including health care professionals who provide such services. Service reviews can be used by health care providers to identify opportunities to improve the quality of care.

Web 2.0 gave rise to the publishing of one's own content and commenting on other user's content on online platforms that provide social networking services. On mainstream social media such as Twitter, Facebook, Instagram, YouTube, and Reddit, patients can organise their fora around groups, hashtags, or influencer users. The primary purpose of these conversations is to exchange information and provide social support online. More specialised websites such as those described in Table 2.6 serve the same purpose. Spontaneous narratives published on these media represent a valuable source for identifying patients' needs, especially the unmet ones.

2.3.2 Data authors

This section provides answers to RQ3 and to RQ4. It discusses the characteristics of those who authored the types of narratives discussed in the previous section. First, the focus is on the roles within health and care of the authors in relation to RQ3, which is followed by their demographic characteristics in relation to RQ4. There are 5 identified roles with respect to health and well-being among the authors of spontaneously generated narratives considered in this review: sufferer, addict, patient, carer, and suicide victim, details can be found in Table 2.7. Some of these roles may overlap, for example, a sufferer or an addict can also be a patient if they are receiving a medical treatment for their medical condition.

Demographic factors refer to socioeconomic characteristics such as age, gender, education level, income level, marital status, occupation, and religion. Most studies involving clinical data summarise the demographics of study participants statistically to illustrate the extent to which its findings can be generalised. The focus on spontaneously generated narratives implies that the corresponding studies could not mandate the collection of demographic factors. Instead, they can only rely on information provided by users in good faith. Different Web platforms may record different demographic factors, which may or may not be accessible to third parties. Nonmandatory user information will typically give rise to missing values. Moreover, demographic information is difficult to verify online, which raises the concerns over the validity of such information even when it is publicly available.

Table 2.8 states which demographic factors, if any, are recorded when a user registers an account on the given online services and which ones are accessible online. Only age and gender are routinely collected, but not necessarily shared publicly. Therefore, it should be noted when SA is used to analyse such data to address a clin-

Role	Description	Studies
Sufferer	A person who is affected by a medical condition.	[182, 255, 151, 219, 67, 65, 135, 189, 47, 279, 62, 270, 169] [208, 282, 61, 174, 48, 106, 113]
Addict	A person who is addicted to a particular substance.	[82, 77, 79, 75, 72]
Patient	A person receiving or registered to receive medical treatment.	[182, 255, 151, 219, 166, 67, 65, 45, 246, 126, 135, 189, 190, 47, 279] [62, 270, 169, 208, 282, 61, 111, 42, 48, 95, 44, 234, 186, 133, 224]
Carer	A family member or friend who regularly looks after a sick or disabled person.	[255, 45, 246, 126, 135, 189, 208, 282, 61]
Suicide victim	A person who has committed suicide.	[43, 73, 90, 100, 158, 173, 176, 188, 200, 204, 216, 218, 229] [230, 248, 252, 268, 271, 273, 276]

Table 2.7: The roles of authors with respect to health and well-being.

ical question, then the findings should be interpreted with caution as it may not be possible to generalise them across the relevant patient population. Out of 109 studies considered in this systematic literature review, only 5 reported the demographics factors [195, 279, 174, 77, 105]. Age was discussed in 3 studies [279, 174, 77], whereas gender was analysed in 3 studies [195, 77, 105].

2.3.3 Areas and applications

This section answers questions RQ5 and RQ6. It discusses the areas of health and well-being covered by the given datasets with reference to RQ5 and it also discusses the practical applications of SA within these areas in relation to RQ6.

Support groups provide patients and carers with practical information and emotional support to cope with health-related problems. The ability to record these conversations online offers an opportunity to study and measure unmet needs of different health communities. These communities tend to be formed around health conditions with high severity and chronicity rates. Not surprisingly, SA has been used to study communities formed around cancer, mental health problems, chronic conditions from asthma to multiple sclerosis, pain associated with these conditions, eating disorders, and addiction, these are shown in Table 2.9. Studying the opinion expressed in spontaneous narratives offers an opportunity to improve health care services by taking into account unforeseen factors. For example, the content of social media can be used to continually monitor the effects of medications after they have been licensed to identify previously unreported adverse reactions [151], or as in [118] to understand why are the rates low of medication adherence. Similarly, SA can be used to differentiate between suicidal and nonsuicidal posts, after which a real-time online counseling can be offered [43].

The provision of health care services itself has been the subject of SA. Table 2.10 outlines different treatments and services discussed by patients whose opinions have been studied by means of SA. Patient reviews of specific medications can support their de-

Platform	Age	Gender	Education level	Income level	Marital status	Occupation	Religion	Used in
Twitter	?/U	?/N	X/N	X/N	X/N	X/N	X/N	see Section 2.3.1
Facebook	M/U	M/U	?/U	X/N	?/U	?/U	?/U	[219, 137]
Instagram	M/U	M/U	X/N	X/N	X/N	X/N	X/N	[217]
YouTube	M/U	?/U	X/N	X/N	X/N	X/N	X/N	[195, 166]
Reddit	X/N	X/N	X/N	X/N	X/N	X/N	X/N	[43, 168, 69]
Amazon	X/N	X/N	X/N	X/N	X/N	X/N	X/N	[40]
RateMDs	X/N	X/N	X/N	X/N	X/N	X/N	X/N	[45, 246, 126]
WebMD	M/U	?/U	X/N	X/N	X/N	X/N	X/N	[255, 135]
Ask a Patient	M/Y	M/Y	X/N	X/N	X/N	X/N	X/N	[189, 190]
DrugLib.com	M/Y	M/Y	X/N	X/N	X/N	X/N	X/N	[255, 189, 190, 47]
Breastcancer.org	M/U	?/U	X/N	X/N	X/N	?/U	X/N	[279, 62]
MedHelp	?/U	M/U	X/N	X/N	X/N	X/N	X/N	[182, 67, 65, 270, 169]
DailyStrength	M/U	M/U	X/N	X/N	X/N	X/N	X/N	[255, 151]
Cancer Survivors Network	?/U	?/U	X/N	X/N	X/N	X/N	X/N	[208, 282, 61]
NHS website	?/U	?/U	?/U	X/N	X/N	X/N	X/N	[111]
DiabetesDaily	?/U	?/U	X/N	X/N	X/N	?/U	X/N	[42]
Drugs.com	M/N	X/N	X/N	X/N	X/N	X/N	X/N	[80]

Table 2.8: Recording and accessing demographic factors. (?: optional recording; X: recording not available; M: recording mandatory; U: user-specific access; N: not accessible online; Y: accessible online; NHS: National Health Service).

Problem	Studied in
cancer	[199, 278, 282, 93, 113], oral [81], lung [169], breast [67, 65, 279, 62, 270, 169, 208, 61, 183], cervical [81], prostate [182], colorectal [57, 208, 61], cancer screening [178]
mental health	[112], depression [137, 217, 139, 113], dementia [196] suicide [43, 73, 90, 100, 158, 173, 176, 188, 200, 204, 216, 218, 229, 230, 248, 252, 268, 271, 273, 276]
chronic condition	[206] diabetes [222, 104, 199, 135, 169, 42, 113, 41], Chron's diesase [219, 67, 65], multiple sclerosis [212, 106], asthma [174]
eating disorder	[247], obesity [142], anorexia [195, 234]
addiction	smoking [77, 79, 75, 72, 168, 69], cannabis [82]
pain	[241, 40], fibromyalgia [114]
infectious disease	Ebola [194, 211], latent infectious disease [160]
quality of life	[201, 64, 70, 224]

Table 2.9: Health-related problems studied by sentiment analysis.

cision making but can also be explored to support shared decision making, ultimately influencing health outcomes and health care utilisation. Patient reviews of health care services can reveal how the services are experienced in practice [275, 45, 246, 126, 111, 95, 44, 214, 186, 128], help improve communication between patients and health care providers, and identify opportunities for service improvement, again influencing health outcomes and health care utilisation. In terms of disease prevention, it is important to understand potential obstacles to population-based intervention approaches such as vaccination [284, 98, 97, 81, 211, 172, 185, 277] and in turn develop strategies for vaccine uptake [277]. Additionally, reviews can support the development of strategies to counter the proliferation of misinformation [172] and analyse the dissemination of information on online platforms [177]. For example, the authors in [185] introduced Crowdbreaks, an open platform which allows tracking of health trends, mostly towards vaccination. Social media data can help identify concerns and ultimately improve health outcomes [41]. Patients' opinions can help health practitioners gain insight into the reasons why some patients may opt for traditional and complementary medicine [93]. Alternatively, understanding patients' experience with different treatments can support creation of personalised therapy plans [278, 106]. Patients' text can in addition serve as an indicator of a technically troubled dialysis [133]. SA can be used to continually monitor online conversations to automatically create alerts for community moderators when additional support is needed [135, 208]. Practical support can be provided by making online health information more accessible [67, 65]. In particular, such information can help carers provide better care to patients [270] and possibly involve computer-aided diagnosis [234].

2.3.4 Methods used for sentiment analysis

This section addresses RQ7 and RQ8. A range of methods along with their implementation that have been used to perform SA are described in this section with respect to RQ7. This is followed by their classification performance to establish the state of the

Treatment	Studied in
Medication	[255, 151, 219, 189, 190, 47, 48, 118, 165, 80]
Vaccine	[284, 98, 81, 211, 172, 185, 277]
Surgery	[164]
Orthodontic	[191, 166]
Physician	[45, 246, 126, 91, 186]
Health care	[275, 85, 111, 95, 44, 214, 186, 128]

Table 2.10: Health care treatments studied by sentiment analysis.

art, in relation to RQ8.

Traditionally, lexicon-based SA methods classify the sentiment as a function of the predefined word polarities [194, 85]. One of such solutions is SentiStrength [29], a lexicon-based SA tool which was used in [160, 166, 106, 104, 247, 224]. Lexicon-based methods are the simplest kind of rule-based methods that focus on the polarity of individual words. In general, rather than focusing on individual words, rule-based methods focus on more complex patterns, typically implemented using regular expressions [158, 176, 188, 204, 229, 230, 248, 276, 70]. Most often, these rules are used to extract features that are relevant to SA, whereas the actual classification is based on machine learning algorithms. One of the SA tools that uses lexicons and simple rules is Valence-Aware Dictionary and Sentiment Reasoner (VADER) [136], it was used in [211, 118, 168, 69, 206, 41, 128, 91]. Table 2.11 provides information about specific machine learning algorithms used. Specific implementations of these algorithms that were used to support experimental evaluation are listed in Table 2.12.

To establish the state of the art of SA in health and well-being the performance of different classification algorithms are summarised in Table 2.13 and Table 2.14. Classification performance measures include accuracy (A), precision (P), recall (R), and F-measure. These measures are calculated using true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), in the following way:

$$A = (TP + TN)/(TP + FP + TN + FN)$$

$$P = TP/(TP + FP)$$

$$R = TP/(TP + FN)$$

$$F - measure = 2 * P * R/(P + R)$$

Although a wide range of methods was used, their performance was rarely systematically tested. According to the no free lunch theorem [259], there is no universally best learning algorithm. In other words, the performance of machine learning algorithms depends not only on a specific computational task at hand, but also on the properties of data that characterise the problem. SVMs proved to be the most popular choice (see Table 2.13), which outperformed naïve Bayes classifier (NB) [82, 98, 67, 268, 164, 136, 113] and random forest [98, 43, 67, 113]. On occasion, it was outperformed by other methods, for example, NB [111, 95], maximum entropy [279], and decision tree [137].

As it can be seen from Table 2.13, accuracy is not routinely reported, which makes it difficult to generalise the findings and compare them with SA performance in other domains. Nonetheless, the accuracy does not fall below 70%. On average, accuracy is around 80%. This is well below accuracy achieved in SA of movie reviews, which is typically well over 90% [272, 127, 110, 138]. However, it is not straightforward to attribute these results to the intrinsic differences between the domains and their respective sublanguages because of the different choices in methods used. The methods tested on movie reviews are based on deep learning, whereas the methods tested on health and well-being related narratives still rely on traditional machine learning with only 6 studies using neural networks [57, 80, 234, 186, 113, 277]. This may be due to the availability of data. Movie reviews are not only publicly available, but also come ready with annotations in the form of star rating. On the other side, health and well-being narratives may contain sensitive information and, therefore, cannot be routinely

Algorithm	Description	Used in
Support vector machine	Builds a classification model as a hyperplane that maximises the margin between the training instances of 2 classes.	[284, 82, 98, 97, 137, 67, 279, 61, 111, 73, 90, 100, 158, 173, 176, 188, 200] [216, 218, 248, 268, 271, 72, 95, 81, 164, 113]
Naïve Bayes classifier	A probabilistic classifier based on Bayes theorem and an assumption that features are mutually independent.	[82, 194, 98, 178, 67, 135, 189, 190, 111, 229, 230, 268] [271, 72, 95, 164, 113]
Maximum entropy	A probabilistic classifier based on the principle of maximum entropy.	[189, 190, 279, 252, 271]
Conditional random fields	A method for labelling and segmenting structured data based on a conditional probability distribution over label sequences given an observation sequence.	[158, 271]
Decision tree learning	A method that uses inductive inference to approximate a discrete-valued target function, which is represented by a decision tree.	[137, 111, 176, 268, 95, 139]
Random forest	An ensemble learning method that fits multiple decision trees on various data samples and combines them to improve accuracy and control overfitting.	[98, 67, 113]
AdaBoost	AdaBoost combines multiple weak classifiers into a strong one by retraining and weighing the classifiers iteratively based on the accuracy achieved.	[279, 208, 282, 61]
k-nearest neighbors	A nonparametric, instance-based learning algorithm based on the labels of the k nearest training instances.	[137, 176]
Logistic regression	A method for modeling the log odds of the dichotomous outcome as a linear combination of the predictor variables.	[82, 61, 273, 139]
Convolutional neural network	A feed-forward neural network that learns to extract salient features that are useful for the given prediction task. Convolutions are used to filter features by using nonlinear functions. Pooling can then be used to reduce the dimensionality.	[57, 113]
Recurrent neural network	Architecture that is designed to process the data sequentially, it contains a recurrent connection (loop) within the architecture.	[234]

Table 2.11: Machine learning algorithms used in sentiment analysis related to health and well-being.

Library	Description	Used in
SVMlight [31]	An implementation of SVMs in C.	[188, 216, 271]
PySVMLight [25]	A Python binding to the SVMlight (see above).	[90]
LIBLINEAR (LIBSVM) [16]	Integrated software for support vector classification, regression, and distribution estimation. It supports multiclass classification.	[98, 61, 73, 100, 158, 173, 200, 248, 83]
Weka [35]	A Java library that implements a collection of machine learning algorithms.	[275, 255, 98, 67, 65, 45, 135, 61, 111, 229, 230, 83, 165]
scikit-learn [27]	A Python library that implements a collection of machine learning algorithms.	[43, 79, 93, 113]
Keras [14]	A high-level neural networks API written in Python.	[278, 186]
TextBlob [32]	A Python library that supports NLP and implements a collection of machine learning algorithms.	[278, 43]
PyTorch [203]	Open source ML library used for computer vision and NLP.	[113]
TensorFlow [39]	Free and open-source software library for machine learning and artificial intelligence.	[80]

Table 2.12: Implementations of machine learning algorithms.

collected en masse. The fact that deep learning does require large amount of data for training may partly explain the preferences toward different types of methods.

Similarly, deep learning is commonly used to support SA of service and product reviews. However, in these domains, the results are closer to those in health and well-being with just over 80% for service reviews and just below 80% for product reviews [265, 132, 155, 71]. The performance still lags behind the state of the art achieved in these 2 domains when measured by F-measure, which was found to be below 60% on average and can go as low as 45%. F-measure achieved on service and product reviews was found to be in between 70s and 80s [265, 266, 156, 249]. In summary, the performance of SA of health and well-being narratives is much poorer than that in other domains, but it is yet unclear if this is because of nature of the domain, the size of training datasets, or the choice of methods. In addition to the choice of methods, their performance largely depends on the choice of features used to represent text. To support basic linguistic preprocessing, most studies used Stanford CoreNLP [175] (e.g., [255, 189, 190, 188, 200, 248, 252, 271, 273, 214]) and Natural Language Toolkit [167] (e.g., [43, 279, 216, 252, 95, 93, 211]). Both libraries represent general purpose NLP tools, which may not be suitable for processing certain sublanguages [103].

Study	**Algorithm**	**Accuracy(%)**	**Precision(%)**	**Recall(%)**	**F-measure(%)**
[81]	SVM	70	-	-	-
[73]	SVM	-	55.72	54.72	55.22
[90]	SVM	-	-	-	53.31
[100]	SVM	-	49	46	47
[158]	SVM + CRG+ rules	-	60.1	36.8	45.6
[173]	SVM	-	51.9	48.59	50.18
[176]	KNN, **DT + SVM + rules**	-	49.92	50.55	50.23
[188]	SVM + rules	-	41.79	55.03	47.5

Continued on next page

Table 2.13 – continued from previous page

Study	Algorithm	Accuracy(%)	Precision(%)	Recall(%)	F-measure(%)
[200]	SVM, **rules**	-	53.8	53.9	53.8
[204]	rules	-	45.98	44.57	45.27
[216]	SVM	-	46	54	49.41
[218]	SVM	-	55.09	48.51	51.59
[229]	NB, rules, **NB + rules**	-	57.09	55.74	56.4
[230]	NB + rules	-	54.96	51.81	53.34
[248]	SVM, **SVM + rules**	-	-	-	50.38
[252]	ME	-	57.89	49.61	53.43
[268]	**SVM + rules**, NB, DT	-	56	62	59
[271]	SVM + NB + ME + CRF + lexicon	-	58.21	64.93	61.39
[273]	LR	-	51.14	47.64	49.33
[111]	SVM, **NB**, DT, bagging	88.6	-	-	59
[135]	NB	-	-	-	54
[208]	AdaBoost	79.2	-	-	-
[279]	SVM, AdaBoost, **ME**	79.4	-	-	-
[282]	AdaBoost	79.2	-	-	-
[189]	NB, ME, **rules**	-	85.25	65	73.76
[190]	NB, **ME**	-	84.52	66.67	74.54
[284]	SVM	88.6	-	-	-
[61]	SVM, LR, **AdaBoost**	79.2	-	-	-
[82]	**SVM**, NB, LR	-	71.47	66.91	67.23

Continued on next page

Table 2.13 – continued from previous page

Study	Algorithm	Accuracy(%)	Precision(%)	Recall(%)	F-measure(%)
[95]	SVM, **NB**, DT	-	-	-	84
[164]	**SVM**, NB	-	63	82	73
[194]	**NB**, lexicon-based	-	75.8	74.3	73
[57]	CNN	76.6	73.7	76.6	73.6
[72]	SVM + NB	82.04	-	-	-
[98]	**SVM**, NB, RF	-	68.73	51.42	58.83
[97]	SVM	-	78.6	78.6	78.6
[139]	LR, **DT**	75	76.1	-	-
[178]	NB	80	-	-	-
[222]	n-gram	-	81.93	81.13	81.24
[67]	**SVM**, NB, RF	-	-	-	82.4
[137]	SVM, KNN, **DT**	-	58	99	73
[234]	lexicon-based, **RNN**	-	69	72	70
[186]	NN	82	83	82	82
[113]	NB, RF, **SVM**, CNN	-	87.82	78.94	83.06
[80]	LSTM, CNN, **BERT + BiLSTM**	-	-	-	90.46
[277]	ELMo+BiGRU, GPT, **BERT**	-	-	-	76.90

Table 2.13: Classification performance. Where multiple algorithms were compared, the performance of the best performing algorithm is indicated by a bold typeset. (-: not applicable).

Aggregated value	Accuracy(%)	Precision(%)	Recall(%)	F-measure(%)
Minimum	70.00	41.79	36.8	45.27
Maximum	88.6	87.82	99	90.46
Median	79.30	58.10	55.74	59
Mean	79.99	63.38	62.03	63.27
Standard deviation	4.96	13.23	14.57	13.29

Table 2.14: Overall classification performance.

2.3.5 Resources

This section answers RQ9. It provides an overview of practical resources that can be used to support development of SA approaches in the context of health and well-being. An overview of lexicons that were utilised in previous studies that have been covered are listed in Table 2.15. Apart from OpinionKB [189], none of the remaining lexicons were developed specifically for applications to health and well-being. To determine how much of their content is specific to health and well-being we cross-referenced against the Unified Medical Language System (UMLS) [58] using MetaMap Lite [88]. This analysis was limited to publicly available lexicons that provide categorical labels of sentiment polarity. The results are shown in Figure 2.2. On average, 18.55%, with standard deviation of 0.0603, of each lexicon accounts for sentimentally polarised UMLS terms. In relative terms, this accounts for a significant portion of each lexicon given their general purpose. In absolute terms, the number of these terms ranges from as little as 330 in WordNet-Affect to as much as 11,687 in SentiWordNet. Knowing that the UMLS currently contains over 11 million distinct terms, we can observe that at most 0.1% of its content is covered by an individual lexicon referenced in Figure 2.2. This means that lexicon-based SA approaches will, by and large, ignore the terminology related to health and well-being.

Resource	Description	Used in
Affective Norms for English Words [60, 1]	A set of normative emotional ratings for a large number of words in terms of pleasure, arousal, and dominance.	[217, 40, 200]
AFINN [187, 2]	A list of 2477 words and phrases manually rated for valence with an integer between -5 (negative) and 5 (positive).	[241, 40, 270]
Harvard General Inquirer [235, 13]	A lexicon attaching syntactic, semantic, and pragmatic information to words. It includes 1915 positive and 2291 negative words.	[67, 65]
LabMT 1.0 [94, 15]	A list 10,222 words, their average happiness evaluations according to users on Mechanical Turk.	[85, 217]
Multi-Perspective Question Answering [257, 18]	A subjectivity lexicon that provides polarity scores for approximately 8000 words.	[151, 188, 248, 75]
Emotion Lexicon (also called EmoLex) [184, 21]	A list of words and their associations with 8 basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and 2 sentiments (negative and positive). The annotations were done manually by crowdsourcing.	[151, 105]
OpinionKB [189, 24]	A knowledge base of indirect opinions about drugs represented by quadruples (e, a, r, p), where e refers to the effective entity, a refers to the affected entity, r is the effect of e on a, and p is the opinion polarity.	[189]
Opinion Lexicon [131, 23]	A list of around 6800 positive and negative opinion words.	[212, 151, 62, 230, 70]
SentiSense [66, 28]	A lexicon attaching emotional category to 2190 WordNet synsets, which cover a total of 5496 words.	[67, 65]
SentiWordNet [51, 30]	An extension of WordNet that associates each synset 3 sentiment scores: positivity, negativity, and objectivity.	[255, 222, 189, 190, 47, 169, 90, 230, 48, 214, 165]
WordNet-Affect [236, 37]	An extension of WordNet that correlates a subset of synsets suitable to represent affective concepts with affective words. Its hierarchical structure was modelled on the WordNet hyponymy relation.	[158, 188, 218, 230]

Table 2.15: Lexical resources for sentiment analysis.

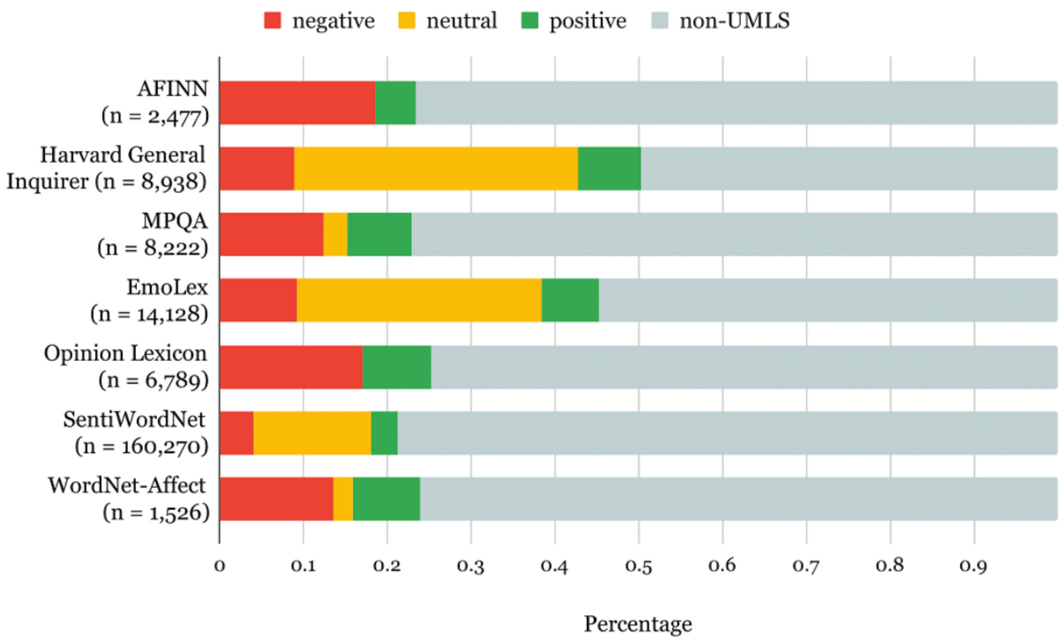

Figure 2.2: The representation of the UMLS in sentiment lexicons.

Extending the UMLS by including sentiment polarity would address this gap, but this problem is nontrivial as lexicon acquisition has been known to be a major bottleneck for SA. Lessons can be learnt from existing research that focuses on automatic acquisition of sentiment lexicons. These approaches can be divided into 2 basic categories: corpus- and thesaurus-based approaches. Corpus-based approaches operate on a hypothesis that words with the same polarity co-occur in a discourse. Therefore, their polarity may be determined from their co-occurrence with the seed words of known polarity [120, 243, 240, 99]. In this context, MEDLINE [20] would be an obvious source for assembling a large corpus. Similarly, thesaurus-based approaches exploit the structure of a thesaurus (e.g., WordNet [180]) to infer polarity of unknown words from their relationships to the seed words of known polarity [145, 141, 119, 96, 170]. They rely on a hypothesis that synonyms (e.g., trauma and injury) have the same polarity, whereas antonyms (e.g., ill and healthy) have the opposite polarity. Starting with the seed words, the network of lexical relationships is crawled to propagate the known polarity in a rule-based approach. The structure of the UMLS could be exploited in a similar manner to infer the sentiment of its terms.

2.4 Discussion

The following section provides the principal findings of the systematic literature review. The topic of this systematic review is SA of spontaneously generated narratives in relation to health and well-being. Specifically, this systematic review was conducted with the aim of answering research questions specified in Table 2.1. It identified a total of 109 relevant studies, which were used to support the findings, which are summarised here.

2.4.1 What Are the Major Sources of Data?

The majority of data were collected from the mainstream social multimedia and Web-based retailing platforms. Mainstream social media provides a generic platform to engage patients. However, their use of social media in the context of disclosing protected health information may raise ethical issues. The need to engage patients online while fully complying with data protection regulations has led to the proliferation of websites and networks developed specifically to provide a safe space for sharing health-related information online. Within this systematic review 11 such platforms were identified, they are listed, along with more details in Table 2.6. In addition to user-generated content, the fifth i2b2/VA/Cincinnati challenge in NLP for clinical data [207] represents an important milestone in SA research related to health and well-being. The corpus used for this shared task contained 1319 written notes left behind by people who died by suicide. This is one of the few datasets that have been made available to the research community. Owing to ethical concerns, the data used in the studies included in this systematic review are usually not released publicly to support further research and evaluation. This makes it difficult to benchmark the performance of SA in health and well-being, and test the portability of methods developed. In addition, the lack of sufficiently large datasets prevents the use of state-of-the-art methods such as deep learning (see Table 2.13).

2.4.2 What Is the Originally Intended Purpose of Spontaneously Generated Narratives?

Web 2.0 gave rise to the self-publishing and commenting on other user's content on online platforms. On mainstream social media such as Twitter, Facebook, Instagram, YouTube, and Reddit, patients can self-organise around groups, hashtags, and influencer users. The primary purpose of these conversations is to exchange information and provide social support online. More specialised websites such as those described in Table 2.6 serve the same purpose.

2.4.3 What Are the Roles of Their Authors Within Health and Care?

There are 5 roles that are identified with respect to health and well-being among the authors of the spontaneously generated narratives considered in this review (see Table 2.7): a sufferer (a person who is affected by a medical condition), an addict (a person who is addicted to a particular substance), a patient (a person receiving or registered to receive medical treatment), a carer (a family member or a friend who regularly looks after a sick or disabled person), and a suicide victim (a person who has committed suicide). Some of these roles may overlap, for example, a sufferer or an addict can also be a patient if they are receiving a medical treatment for their medical condition.

2.4.4 What Are Their Demographic Characteristics?

The focus on spontaneously generated narratives implies that the corresponding studies could not mandate the collection of demographic factors. Different Web platforms may record different demographic factors, which may not be accessible to third parties. Demographic information is also difficult to verify online, which raises the concerns over the validity of such information even when it is publicly available. Table 2.8 states which demographic factors, if any, are recorded when a user registers an account on the given online services and which ones are accessible online. Only age and gender are routinely collected, but not necessarily shared publicly. Therefore, any findings resulting from these data should be interpreted with caution as it may not be possible to generalise them across the relevant patient population. Out of 109 studies considered in this review, only 5 reported the demographic characteristics.

2.4.5 What Areas of Health and Well-Being Are Discussed?

Online communities tend to be formed around health conditions with high severity and chronicity rates. Not surprisingly, SA has been used to study communities formed around cancer, mental health problems, chronic conditions from asthma to multiple sclerosis, pain associated with these conditions, eating disorders, and addiction, these are listed in Table 2.9. The provision of health care services itself has been the subject of SA. Different treatments and services discussed by patients whose opinions have been studied by means of SA include medications, vaccination, surgery, orthodontic services, individual physicians, and health care services in general.

2.4.6 What Are the Practical Applications of Sentiment Analysis?

Analysing the sentiment expressed in spontaneous narratives offers an opportunity to improve health care services by taking into account unforeseen factors. For example, social media can be used to continually monitor the effects of medications to identify previously unknown adverse reactions. Similarly, SA can be used to differentiate between suicidal and nonsuicidal posts, after which a real-time online counseling can be offered. Patient reviews of specific medications can support their decision making but can also be explored to support shared decision making, ultimately influencing health outcomes and health care utilisation. Patient reviews of health care services can help identify opportunities for service improvement, thus influencing health outcomes and health care utilisation. In terms of disease prevention, patients' opinions can help health practitioners understand potential obstacles to population-based intervention approaches such as vaccination. Understanding patients' experience with different treatments can support creation of personalised therapy plans.

2.4.7 What Methods Have Been Used to Perform Sentiment Analysis?

A wide range of methods have been used to perform SA. Most common choices include SVMs, naïve Bayesian learning, decision trees, logistic regression, and adaptive boosting. Other approaches include maximum entropy, conditional random fields, random forests, and k-nearest neighbors. The findings show strong bias toward traditional machine learning. Only 6 studies used deep learning to perform SA, out of which 5 of them were additionally included in the review as a result of the update. This is in stark contrast with general trends in SA research.

2.4.8 What Is the State-of-the-Art Performance of Sentiment Analysis?

On average, accuracy is around 80%, and it does not fall below 70%. This is well below accuracy achieved in SA of movie reviews, which is typically well over 90%. In SA of service and product reviews, the results are closer to those in health and well-being with on average around 80%. However, the performance still lags behind the state of the art achieved in these 2 domains when measured by F-measure, which was found to be below 60% on average. F-measure achieved on service and product reviews is found to be above 70% and 80%, respectively. In summary, the performance of SA of health narratives is much poorer than that in other domains.

2.4.9 What Resources Are Available to Support Sentiment Analysis Related to Health and Well-Being?

A wide range of lexicons were utilised in studies covered by this systematic review, they are listed in Table 2.15. Notably, out of 11 lexicons, only 1 was developed specifically for a domain related to health and well-being. The lack of domain-specific lexicons may partly explain the poorer performance recorded in this domain.

2.5 Summary

In summary, the systematic literature review described in this Chapter has uncovered multiple opportunities to advance research in SA related to health and well-being. Keeping in mind the no free lunch theorem, researchers in this area need to put more effort in systematically exploring a wide range of methods and testing their performance. Community efforts to create and share a large, anonymized dataset would enable not only rigorous benchmarking of existing methods but also exploration of new approaches including deep learning. This should help the field catch up with the most recent developments in SA. The creation of domain-specific sentiment lexicons stands to further improve the performance of SA related to health and well-being. Although many studies have dealt with automatic construction of domain-specific sentiment lexicons using methods such as random walks, no such studies have been identified in this systematic review. Finally, health-related applications of SA require systematic collection of demographic data to illustrate the extent to which the findings can be generalised.

Chapter 3

Measuring the Performance of Sentiment Analysis Tools

The work presented in this Chapter has been accepted for the HealTAC conference 2020. It is based on the following paper: Žunić Anastazia, Corcoran Padraig, and Spasić Irena. Improving the performance of sentiment analysis in health and wellbeing using domain knowledge. *HealTAC 2020: Healthcare Text Analytics Conference*, London, UK, Online, 2020.

3.1 Introduction

The systematic literature review described in Chapter 2 revealed that sentiment analysis (SA) has poorer performance in health and well-being compared to the other domains. In this Chapter, we provide new evidence of this fact by evaluating five publicly available SA tools, along with the ensemble method that combines them, on two different domains. The performance of SA tools is evaluated on top of two publicly available datasets. These datasets represent user-generated content from the online platforms. One of these, namely drug reviews, is related to health and well-being. The second one, movie reviews, is used for cross-domain comparison of SA. First, we compare the performance of five general-domain SA tools. Furthermore, we investigate how their performance changes on health and well-being related content by accounting for the bias towards the negative sentiment using domain knowledge. This was done by semantically enriching the data using explicit domain knowledge which is formally modelled by the Unified Medical Language System (UMLS) [58].

As mentioned before, this thesis is focused on the domain of health and well-being, where the content is generated by the users, patients and health care professionals alike. One of the biggest motivators for exploring SA in this domain is the potential for deeper understanding of patients' needs, including the unmet ones. It may give a health care professional an insight into patients' opinions on treatments or services that they provide. For example, negative reviews can be used by clinicians and hospitals to identify services that need to be improved. In the pharmaceutical industry it can identify side effects of medications when they enter postmarketing surveillance. Based on the sublanguage theory of Zellig Harris [117], we hypothesise the following:

- (H1) the sublanguage of the user-generated content varies across domains,
- (H2) the sublanguage of health and well-being is biased towards negative sentiment,
- (H3) medical knowledge can be used to address this bias and consequently improve the performance of SA.

This Chapter is organised as follows. Section 3.2 provides the overview of SA tools that are used for the purpose of comparison. Section 3.3 provides the details about the datasets that have been utilised for the evaluation of the SA tools. A concept abstraction approach as a way of accounting for the bias towards negative sentiment associated with medical concepts is proposed in Section 3.4. Section 3.5 provides the evaluation results. Summary is given in Section 3.6.

3.2 Sentiment analysis tools

Five off-the-shelf general-domain SA tools that have been used for the comparison of SA across two different domains, one being health and well-being related and the other one related to movie reviews, are:

- SentiStrength - a lexicon-based method designed to analyse the sentiment of short texts such as those found on social media. It assesses the strength of both negative and positive sentiments on the scales from -1 to -5 and 1 to 5 respectively [29]. Positive and negative sentiment scores are aggregated into a single value between -4 and 4 by simply adding them up.
- TextBlob - utilises a naïve Bayesian approach trained on a corpus of movie reviews to return a polarity score within the range -1 to 1 [32].
- NLTK - lexicon - uses an opinion lexicon described in [131, 23] to calculate the sentiment score by aggregating the polarity scores of individual words to output a value from “negative”, “neutral” or “positive” [167].
- NLTK - VADER - supports SA of social media. It combines a lexicon with rules to obtain a sentiment score ranging from -1 to 1 [136].
- Stanford CoreNLP - uses recursive neural networks to classify the sentiment into the values within a range from 0 to 4, where the values 0 and 4 correspond to extreme negative and positive sentiments, respectively. It was trained on a set of parse trees, whose subtrees were annotated with sentiment values in order to enable fine-grained SA [228].

The SA experiments were performed at the document-level. To enable direct comparison, all outputs were normalised to a 3-point Likert scale: negative - neutral - positive, this is shown in Table 3.1.

SentiStrength	TextBlob	NLTK-lexicon	NLTK-VADER	Stanford CoreNLP	Label
-4, -3, -2	[-1, -0.2)	negative	[-1, -0.2)	[0, 1.6)	negative
-1, 0, 1	[-0.2, 0.2]	neutral	[-0.2, 0.2]	[1.6, 2.4]	neutral
2, 3, 4	(0.2, 1]	positive	(0.2, 1]	(2.4, 4]	positive

Table 3.1: Normalisation of SA outputs to a single scale.

3.3 Data

For the purpose of the analysis two datasets have been used. These datasets consist of user-generated content, one related to health and well-being (Drugs.com reviews [109]) and the second one related to movies (Amazon Movie reviews [121]). The content of Drugs.com reviews, are namely drug reviews that are generated by the users who used specific medication, whereas the Amazon Movie reviews were generated by the users who have seen a particular movie. To normalise the gold standard labels across all datasets and make them compatible with the 3-point Likert scale, previously mentioned, the star rating attached to the review is used and mapped onto the same scale: negative - neutral - positive (see Table 3.2) following an approach previously employed in [109].

Drugs.com	Amazon Movie	Label
1, 2, 3, 4	1, 2	negative
5, 6	3	neutral
7, 8, 9, 10	4, 5	positive

Table 3.2: Normalisation of gold standard annotations.

From each dataset, a sample of 300 reviews were randomly selected from each class (negative, neutral and positive), yielding a total of 900 reviews. The samples were evenly distributed across the star ratings as well, i.e. 300 negative samples for Amazon Movie dataset were selected in the following way: 150 reviews were randomly selected from Amazon Movie 1-star rating and 150 random reviews from Amazon Movie 2-star rating. Average number of sentences per review over three different classes is

	Drugs.com	**Amazon Movie**
negative	5.61	5.98
neutral	5.56	7.86
positive	5.91	6.83
mean	5.69	6.89
standard deviation	0.15	0.77

Table 3.3: Average number of sentences over sentiment labels.

approximately the same within the health and well-being dataset, whereas the standard deviation in Amazon Movie reviews is a bit larger. In Drugs.com reviews, the average number of sentences per review is 5.7 and in Amazon Movie is 6.9, this can be seen in Table 3.3.

3.4 Abstraction of medical concepts

Hypothesis H1 states that sublanguage related to health and well-being differs most notably with respect to its vocabulary. Additionally, hypothesis H2 claims that the domain-specific vocabulary related to health and well-being is biased toward negative sentiment. Hypothesis H3 claims that medical knowledge can be used to mitigate the bias and consequently improve the performance of SA. To test the hypotheses, H1, H2, and H3, experiments involving such vocabulary have been conducted. Specifically, words were abstracted from medical vocabulary using a domain-specific dictionary, e.g. "Reduced my **pain** by 80% and lets me live a normal life again." would be mapped to "Reduced my **symptom** by 80% and lets me live a normal life again." By abstracting specific concepts, we effectively neutralise the sentiment. In order to support the abstraction of concepts the Unified Medical Language Systems (UMLS) [58], a system that integrates multiple terminologies, classifications and coding standards, was employed. MetaMap Lite [88], a dictionary lookup program, was used to find mentions of concepts from the following semantic types: sign or symptom (SOSY), disease or

Semantic type	Concepts	Examples
SOSY	294	pain, tremble, weak, nausea, numbness, vomiting, illness
DSYN	305	migraine, headache, asthma, hypertension, infection, virus, hepatitis
PATF	71	complication, sensitive, infection, choke, allergies, inflammation, shock
INPO	71	wound, bruises, burns, crushed, injury, bite, fracture
NEOP	58	cancer, tumor, pancreas, heart, stomach, retina, metastasis

Table 3.4: The number of unique concepts matched.

	Drugs.com			Amazon Movie		
Semantic type	**positive**	**neutral**	**negative**	**positive**	**neutral**	**negative**
SOSY	450	586	552	74	101	54
DSYN	400	398	376	477	581	399
PATF	156	154	128	31	33	18
INPO	40	46	46	29	55	27
NEOP	70	78	78	40	74	26

Table 3.5: Distribution of medical concepts across datasets and sentiments.

syndrome (DSYN), pathological function (PATF), injury or poisoning (INPO) and neoplastic process (NEOP). Their choice was based on the most common UMLS semantic types of unique concepts occurring in the corpus of Drugs.com reviews. The number of unique concepts matched from the five semantic types is given in Table 3.4, along with the examples of concepts that belong to these semantic types. Their distribution across the two datasets and sentiment categories is given in Table 3.5. Afterwards, all concept mentions found were then replaced in text by the corresponding semantic type. If a concept was mapped to multiple semantic types the most general semantic type was selected. For example, text such as "**Allergic reaction** after 4 pills. **Hives**, **flushing**, **headache**." would be translated to "**PATF** after 4 pills. **DSYN**, **SOSY**, **SOSY**.".

The distribution of medical concepts across datasets (see Table 3.5) suggests that the sublanguage of the user-generated content varies across domains, in this case the domains being health and well-being and the movie reviews. This shows that the hypo-

Drugs.com	SentiStrength	TextBlob	NLTK-lexicon	NLTK-VADER	Stanford CoreNLP	ensemble
Precision	0.40	**0.48**	0.39	0.43	0.44	0.45
Recall	0.38	0.40	0.39	**0.43**	0.37	0.41
F-measure	0.33	0.36	0.36	**0.40**	0.26	0.38

Table 3.6: SA results on Drugs.com reviews.

Amazon Movie	SentiStrength	TextBlob	NLTK-lexicon	NLTK-VADER	Stanford CoreNLP	ensemble
Precision	0.55	**0.58**	0.41	0.38	0.52	**0.61**
Recall	0.45	0.46	0.45	0.43	**0.53**	**0.56**
F-measure	0.42	0.43	0.40	0.36	**0.48**	**0.57**

Table 3.7: SA results on Amazon Movie reviews.

thesis H1 is true.

3.5 Results

To establish the baseline, the five SA tools were run over the two datasets, the results are shown in Table 3.6 and Table 3.7. In addition, the performance of an ensemble method is reported. The ensemble method sums up the sentiment scores obtained by the underlying SA tools and stratifies the values into three classes: negative (-5 ≤ score ≤ -2), neutral (-1 ≤ score ≤ 1) and positive (2 ≤ score ≤ 5). All scores were macro-averaged as the classes were evenly balanced.

Having performed concept abstraction as described in the previous section, the SA tools were used to analyse the newly processed data. Comparing the results before and after the abstraction on Drugs.com reviews (Table 3.6 and Table 3.8) a slight improvement in SA can be identified when the Drugs.com reviews were mapped to a single UMLS semantic type (SOSY, DSYN, PATF, INPO, or NEOP). The F-measure of the ensemble method increased by 2 percent points when Drugs.com reviews were mapped to DSYN, PATF, INPO and NEOP, whereas an increase of 5 percent points occurred when the reviews were mapped to SOSY. However, a significant boost in the perform-

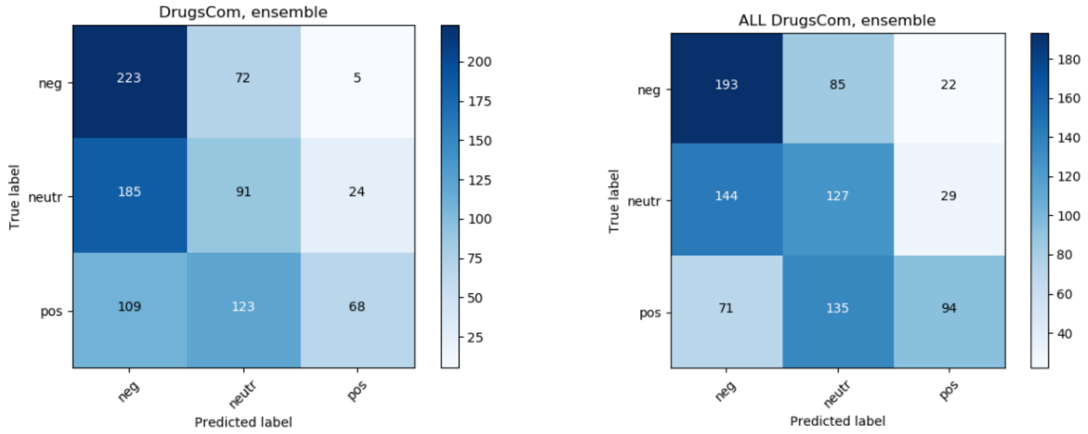

Figure 3.1: Confusion matrices for the ensemble method on Drugs.com reviews, before (left) and after (right) abstraction.

ance has occurred when the abstraction of all five selected UMLS semantic types was performed (see Table 3.9). The F-measure of the NLTK- lexicon method increased by 7 percent points. The ensemble method had the F-measure of 0.38, which increased to 0.45 after the abstraction.

Additionally, from the confusion matrices shown in Figure 3.1 we can observe that the predicted sentiment is less biased towards the negative end following concept abstraction. This suggests that the abstraction of medical concepts does boost the performance of existing SA tools. Consequently, this shows that our hypothesis H3 is correct and that medical knowledge can be used to mitigate the bias towards the negative end and improve the performance of SA.

In order to examine the effect of concept abstraction on SA results, precision, recall and F-measure were measured against each sentiment independently for the Drugs.com reviews before and after concept abstraction (see Table 3.10). The recall over positive sentiment is significantly lower than the recall over negative sentiment, which means that the number of correctly classified positive reviews is low, whereas that is not the case with the negative sentiment. This can also be seen in Figure 3.1. However, precision for the positive sentiment is three times greater before and more than two times

SOSY **Drugs.com**	**SentiStrength**	**TextBlob**	**NLTK-lexicon**	**NLTK-VADER**	**Stanford CoreNLP**	**ensemble**
Precision	**0.42**	**0.49**	0.39	0.42	0.42	**0.46**
Recall	0.38	0.40	**0.40**	**0.44**	0.36	**0.44**
F-measure	**0.34**	0.36	**0.39**	**0.41**	**0.28**	**0.43**

DSYN **Drugs.com**	**SentiStrength**	**TextBlob**	**NLTK-lexicon**	**NLTK-VADER**	**Stanford CoreNLP**	**ensemble**
Precision	**0.41**	**0.49**	0.38	0.43	0.39	0.45
Recall	0.38	**0.41**	0.39	**0.44**	0.36	0.41
F-measure	0.33	**0.37**	**0.37**	**0.41**	**0.27**	**0.40**

PATF **Drugs.com**	**SentiStength**	**TextBlob**	**NLTK-lexicon**	**NLTK-VADER**	**Stanford CoreNLP**	**ensemble**
Precision	0.40	0.48	0.38	0.43	0.43	**0.46**
Recall	0.38	0.40	0.39	**0.44**	0.36	**0.42**
F-measure	0.33	0.36	0.36	**0.41**	**0.29**	**0.40**

INPO **Drugs.com**	**SentiStrength**	**TextBlob**	**NLTK-lexicon**	**NLTK-VADER**	**Stanford CoreNLP**	**ensemble**
Precision	**0.41**	0.48	0.38	0.43	0.44	**0.46**
Recall	0.38	0.40	0.39	**0.44**	0.37	**0.42**
F-measure	0.33	0.36	0.36	**0.41**	**0.30**	**0.40**

NEOP **Drugs.com**	**SentiStrength**	**TextBlob**	**NLTK-lexicon**	**NLTK-VADER**	**Stanford CoreNLP**	**ensemble**
Precision	**0.41**	**0.49**	0.39	**0.44**	0.44	**0.46**
Recall	0.38	0.40	0.39	**0.44**	0.36	**0.42**
F-measure	0.33	0.36	0.36	**0.41**	**0.29**	**0.40**

Table 3.8: SA results on Drugs.com reviews after concept abstraction over single UMLS semantic types.

ALL **Drugs.com**	**SentiStrength**	**TextBlob**	**NLTK-lexicon**	**NLTK-VADER**	**Stanford CoreNLP**	**ensemble**
Precision	**0.42**	**0.49**	**0.44**	0.42	0.38	**0.50**
Recall	0.37	**0.41**	**0.46**	**0.44**	0.35	**0.46**
F-measure	**0.34**	0.36	**0.43**	**0.41**	0.26	**0.45**

Table 3.9: SA results on Drugs.com reviews after the concept abstraction.

greater after the abstraction from the recall. That implies that out of all the reviews that are classified as positives a great number of them really are positive. Consequently,

	negative			neutral			positive		
	P	R	F	P	R	F	P	R	F
before abstraction	0.43	0.74	0.52	0.33	0.33	0.33	0.60	0.19	0.29
after abstraction	0.47	0.64	0.55	0.37	0.42	0.39	0.65	0.31	0.42

Table 3.10: Precision (P), recall (R) and F-measure (F) measure against each sentiment label before and after concept abstraction.

negative sentiment has lower precision than recall because the sentiment is biased towards the negative end, therefore a lot of positive and neutral reviews are classified as negative ones. This shows that our hypothesis H2 is correct and that the sublanguage of health and well-being is biased towards the negative sentiment.

3.6 Summary

In this Chapter we showed that general-domain SA tools do have poorer performance on the data that contains health and well-being related content, which was also a finding from the systematic literature review described in Chapter 2. Furthermore, we showed that the sublanguage of user-generated content varies across domains and that the sublanguage of health and well-being is skewed towards the negative sentiment. Additionally, we demonstrated that the abstraction of domain-specific content related to health and well-being can effectively mitigate the inherent bias towards the negative sentiments and consequently improve the performance of general-purpose SA. In fact, the experiments demonstrated an improvement in F-measure by 7 percent points.

Chapter 4

An Overview of Deep Learning

This project aims to develop a machine learning approach that can analyse interactions of an aspect with the surrounding words in order to infer its sentiment. Deep learning has been singled out as one class of machine learning algorithms that have the capability of learning complex non-linear binary classifications and as such demonstrate the greatest potential for dealing with the problem of aspect-based sentiment analysis. This Chapter provides an introduction into the fundamental concepts of deep learning methods as well as their successful applications to sentiment analysis.

4.1 Introduction

Deep learning algorithms involve the use of large multi-layer neural network models. The use of neural networks in NLP has emerged in the past decade. One of the greatest breakthroughs in NLP happened in the early 2010s with the use of neural networks to learn vector representations of words commonly known as word embeddings. Since then, word embeddings have become fundamental in the application of deep learning to a variety of NLP problems. Previously, words were usually represented with one-hot encoding, which resulted in very large and sparse vectors. On the other hand, the dimension of these vectors are typically much lower than one-hot encoding vectors. Word embeddings are dense vector representations of words that encode the meaning of these words by grouping similar (or related) words close together in the vector space. Word embeddings are learned directly from raw data in a self-supervised manner without any human annotation required. One of the main advantages of word embeddings, relative to one-hot encoding, is that similar words will tend to have similar representations. For example, the pairs of words cat and dog will have more similar vectors than the pair of words cat and hello. On the other hand, in the case of one-hot encoding, all vectors are equally similar to one another. In addition to word embeddings capturing semantic similarity, directions in the embedding space can capture relations between words such as analogies. An example of how the meaning is encoded in word embeddings and how the direction captures relations can be seen in Figure 4.1. Here an analogy between the data points can be described with the equation $king - men + woman = queen$ thus capturing a gender relationship. Similar to the previous example is the analogy of capital relationship where the following equation applies $Rome - Italy + France = Paris$. One of the first and most famous word embedding models is *word2vec*, developed in 2013 by Mikolov et al. in [179]. Another very successful model, albeit a log-bilinear and not a neural one, is *Global Vectors for Word Representation* (*GloVe*) introduced in 2014 by Socher et al. in [205].

The issue with the aforementioned embeddings is that they are static in the sense that each word maps to a constant single vector. This is despite the fact that a single word can take on multiple meanings depending on the context within which it appears. For example, consider the word bank which can mean a financial bank or a river bank where this difference in meaning is a function of the context within which the word appears. Despite this, when using the above embedding methods, the word bank will always have the same constant representation. This problem was addressed by the introduction of contextualised embeddings, which will be described in this and subsequent Chapters in the thesis. The successful application of deep learning to NLP can be attributed in part to the development of word embeddings and later contextualised word embeddings methods.

Neural networks have been extensively applied in various NLP tasks over the past dec-

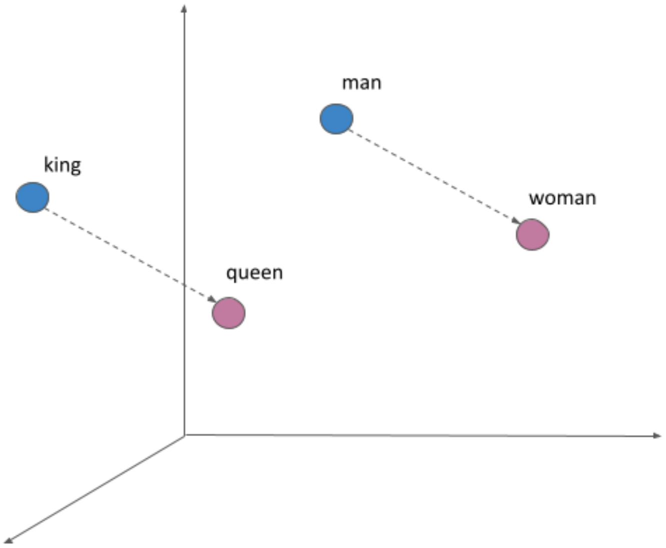

Figure 4.1: The illustration of an analogy between word embeddings in a vector space.

ade, including the successful application to sentiment analysis (SA). In the following sections we describe the fundamental concepts of deep learning methods. In doing so, we describe some of the most common neural network architectures used in NLP.

4.2 Neural networks

The term *neural networks* refers to a collection of algorithms that are designed to recognise patterns within data, where this data can be in the form of images, text, and in general any real-world data. A neural network contains an input layer, a hidden layer and an output layer, see Figure 4.2. A deep neural network is any network that contains more than one hidden layer. The depth of a neural network is defined by the number of layers it contains. A hidden layer in a neural network generally corresponds to a non-linear transformation or mapping. For example, a specific hidden layer commonly used in practice performs a sum of weighted inputs, where a bias is added to this sum,

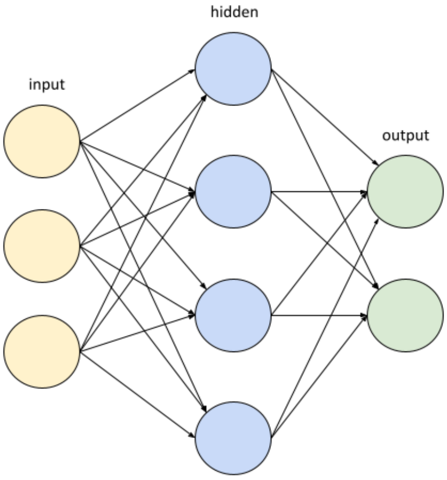

Figure 4.2: The illustration of a neural network.

after which the result is passed to an activation function. An activation function adds non-linearity to a neural network; it decides if the neuron within the network will be activated or not and to what degree. Some of the examples of activation functions are ReLU, sigmoid, hyperbolic tangent (tanh). The choice of an output layer will vary as a function of a given task. For example, in the context of classification, a commonly used output layer is the softmax function. The softmax function calculates the probability distribution over the n classes, i.e. it is a function that turns a vector of N real values into a vector of N real values that sums up to 1. It is calculated using the Equation 4.1, where z_i is the i-th output of the last hidden layer and N is the number of classes.

$$softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}}. \tag{4.1}$$

To measure the correctness of the model, neural networks use a loss function. Broadly speaking, a loss function measures the difference between the actual output and the desired output of the neural network. In practice, the goal is to minimise the value of the loss function and in turn maximise the accuracy of the model. This is achieved with the help of optimization algorithms such as gradient descent. An optimization algorithm iteratively adjusts the parameters or weights of a neural network such that the loss function is minimised. The amount by which the weights of a neural network are updated at each step of the training process is defined by a hyperparameter known as the learning rate.

Most optimization algorithms used to support deep learning are gradient based methods. Examples of such algorithms include gradient descent, stochastic gradient descent and Adam (derived from adaptive moment estimation). Gradient descent is an algorithm that finds a local minimum of a differentiable function and as a result it minimises the loss function. Its main drawback is that it operates on the whole dataset and consequently is more likely to get trapped in a local minimum. On the other hand, stochastic gradient descent operates on random subsets of the dataset and this randomness in the process can help it escape local minima. However, further in the thesis we employ Adam optimizer [147], which is a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments of the gradients.

A common problem that occurs in deep learning is overfitting. This means that the model is trained too much on the training data and consequently will not perform well on the unseen data. One of the ways to prevent the neural network from overfitting is to use dropout. Dropout is a parameter that indicates that a neuron will be turned off during the training process with some probability p. As a consequence, a neural network will not assign high weights to certain features as they might disappear, resulting in the spread of weights across all features.

A dataset is typically split into three parts, training, validation (development) and test dataset. As the names suggest, the model is trained on the training dataset, while the hyperparameters are tuned using the validation dataset. Finally, the trained model is evaluated on the test dataset, which contains the data that has not been previously seen by the model during the training process. An epoch corresponds to one iteration of the training process over the entire training dataset. Sometimes models are trained for hundreds of epochs. However, it is a common practice to train it for several epochs, for example 10. Models are trained for more epochs in order to reduce the loss. The trained model is evaluated using evaluation measures that include precision, recall, F-measure and accuracy. As well as these measures, the value of the loss function is also one of the dominant indicators of how good the model is. Precision measures the number of retrieved documents that are relevant, whereas recall measures the number of relevant documents that are retrieved. They are calculated using true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), as presented in Equation 4.2.

$$Precision = \frac{TP}{TP + FP}, \qquad Recall = \frac{TP}{TP + FN}. \tag{4.2}$$

Additionally, F-measure is a weighted metric which represents a harmonic mean of the aforementioned precision and recall, the equation for its calculation is given in Equation 4.3.

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

The accuracy simply represents the percentage of the correctly classified documents. It is calculated using the Equation 4.4.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4.4}$$

On the other hand, the loss function can be defined in different ways. When performing learning, the objective is to minimise the loss function. The loss function that has been extensively applied in classification tasks is the cross-entropy loss function, it is calculated using the Equation 4.5:

$$loss = -\sum_{n=1}^{N} t_i \log(p_i), \tag{4.5}$$

where t_i is the true label, and p_i is the predicted probability for the i^{th} class. Cross-entropy loss will increase if the softmax predicted probability diverges from the actual label.

4.3 Architectures

Two types of neural network architectures will be discussed within this section; specifically, feedforward neural networks and recurrent neural networks. Feedforward neural networks are models where the information flows forwards only and there are no feedback connections where the output of one layer is fed back to a previous layer. The models that do have the feedback connections are called recurrent networks. More detailed descriptions of these neural network architectures will be given in the remainder of this section.

4.3.1 Recurrent neural networks

The recurrent neural network (RNN) [221] architecture is designed to process data sequentially. It is a feedforward neural network which contains a recurrent connection (loop) within the architecture as shown in Figure 4.3. RNNs are specialized to process a sequence of values such as text. Therefore, they have been widely used in NLP for a number of tasks. The hidden state of the current element represents the memory of the network at the particular time step by capturing information about all previous time steps. The main disadvantage of simple RNN networks is the vanishing gradient problem [123]. Namely, during training each weight receives an update that is proportional to the partial derivative of the error function with respect to its current value. If the gradient becomes vanishingly small, its update will eventually become equal to zero, thus effectively preventing the weight from getting updated. This means that the model will not be able to learn long-term dependencies.

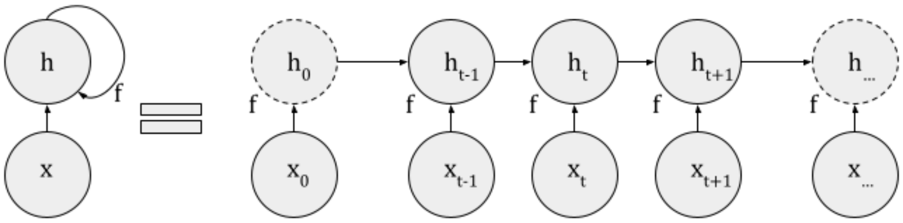

Figure 4.3: The illustration of RNN on the left, and the unfolded RNN on the right, each node is associated with a time instance.

4.3.1.1 Long short-term memory

The long short-term memory (LSTM) architecture is a type of RNN, which improves the modelling of long-range dependencies and overcomes the vanishing gradient problem. It was invented in 1997 by Hochreiter and Schmidhuber in [124]. This architecture contains three types of gates (input, forget and output gates), which are used to calculate the hidden state. The architecture of the LSTM cell is shown in Figure 4.4. The output at the current time step t is calculated using the following set of equations:

$$\begin{aligned} f_t &= \sigma(x_t U^f + h_{t-1} W^f) \\ i_t &= \sigma(x_t U^i + h_{t-1} W^i) \\ o_t &= \sigma(x_t U^o + h_{t-1} W^o) \\ c_t &= f_t \cdot c_{t-1} + i_t \cdot tanh(W_c x_t + U_c h_{t-1} + b_c) \\ h_t &= o_t \cdot tanh(c_t) \end{aligned} \tag{4.6}$$

where x_t is the input embedding, c_t is the cell state, h_t is the hidden state which is the output of the current timestep, $\cdot$ represents the element-wise product, σ is the sigmoid function, and f_t, i_t and o_t are the activation vectors for forget, input, and output gates respectively. With these three gates, the LSTM cell is able to control the flow of information in the following way: the *forget gate* decides what will be kept and what will be removed from the memory, the *input gate* quantifies which information is im-

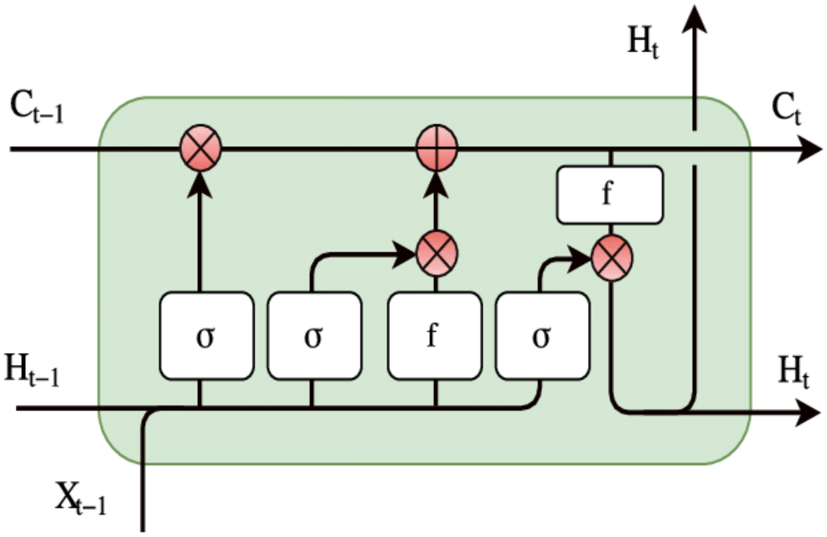

Figure 4.4: Block diagram of LSTM cell. Image taken from [197].

portant and the *output gate* decides what the next hidden state will be. Bidirectional LSTM (BiLSTM) contains two LSTMs, where the information is propagated in both directions, forwards and backwards.

4.3.1.2 Gated recurrent unit

The gated recurrent unit (GRU) [74] has a similar architecture as LSTM, yet it is a bit simpler. It contains two gates, reset and update gates, which are used to memorise the relevant information. The *reset gate* decides how much of the information will be kept, whilst the *update gate* combines the input and forget gate and decides what information to throw away and what information to add.

4.3.2 Convolutional neural networks

Although originally mostly applied in computer vision, convolutional neural networks (CNNs) [153] have recently found many applications in NLP as well. A CNN is a type of the feedforward neural network. The data that is the input to the CNN is

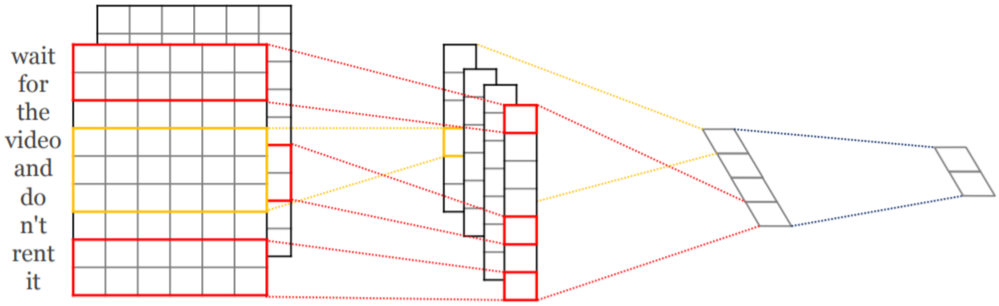

Figure 4.5: CNN for text classification, where different convolutional filters have been applied to the text. Image taken from [146].

processed at once in a single step, unlike RNNs which process the data sequentially. When performing convolution over textual data, a one-dimensional convolution filter is applied over the one-dimensional sentence. On the other hand, in image processing, a two-dimensional filter is applied over the x and y axis. An illustration of convolution over the textual data is shown in Figure 4.5.

4.3.2.1 Graph convolutional networks

Graph convolutional networks (GCN) [148], as the name indicates, performs convolution over a graph. GCNs have been applied to study social networks, biology, neural science, generally in domains where the data is represented in the form of graphs. Likewise, text can also be represented in the form of graphs. In fact, it is one of the natural ways to represent sentences, where vertices correspond to words and edges correspond to relationships between the words. For example, a sentence can be represented as a sequence graph, or as a graph that is extracted from the dependency parse tree of the sentence. There are multiple GCN algorithms, more details will be given in the next Chapter (see Section 5.2.2), where we employ GCN to perform aspect-based SA.

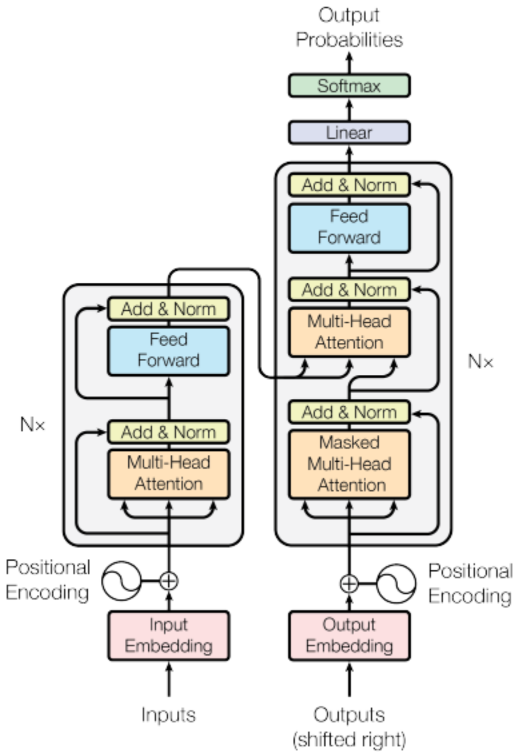

Figure 4.6: Transformer architecture. Image taken from [245].

4.3.3 Transformers

Until 2017, RNNs were the most frequently used NN architectures in NLP. However, since then NLP has been dominated by the transformer-based architecture. Transformers were introduced by Vaswani et al. in [245] in 2017. It is an encoder-decoder neural network architecture that uses a mechanism called neural attention. The main role of the attention mechanism is to determine which words are important for a given task within a given context. Transformers process data in parallel rather than sequentially as it is the case with RNNs. They can be interpreted as fully connected graphs which enables the model to learn the context and to reference words that are further apart in a sentence. The transformer architecture is shown in Figure 4.6. The most popular examples of transformer-based architectures include Bidirectional Encoder Representations from Transformers (BERT) [92], XLNet [272] and Generative Pre-trained Transformer (GPT) [210] with BERT being the most popular choice in research applic-

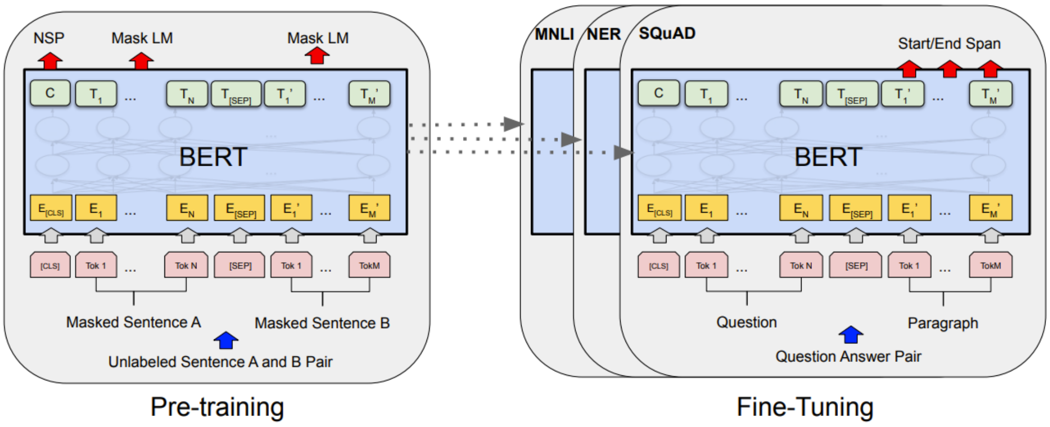

Figure 4.7: BERT architecture. Image taken from [92].

ations due to its open-source licence and the ability to easily fine-tune it for specific NLP tasks.

BERT uses transformer encoders to learn contextual relations between the words, or sub-words, in a bidirectional way, and therefore provides contextualised embeddings. BERT is pre-trained on two tasks simultaneously, masked language modelling (MLM) and next sentence prediction (NSP), see Figure 4.7. MLM is performed by masking a certain percentage of words with a special token [MASK] forcing the model to predict the masked tokens instead. In the NSP task, the model receives a pair of sentences and then it predicts whether or not the second sentence succeeds the first. Being trained in a bidirectional way, BERT represents a language model that can distinguish between the words that are written in the same way but have different meanings depending on their context. In this way, the word $bank$ will have a different representation in the two following sentences: “I went to the $bank$ to open an account.” (financial) and “I went to the $bank$ of the river to get some air” (geographical). During fine-tuning, BERT is initialised with the parameters from the pre-trained model after which these parameters are optimised according to the downstream task using an additional dataset to support the training for that task. Further details on BERT will be provided in Chapter 6 where we describe how we fine-tuned this language model to perform the task of aspect-based SA.

4.4 Neural networks for sentiment analysis

In this section we provide a high-level overview of applications of the aforementioned deep learning architectures in SA, including both general SA and aspect-based SA.

A recent systematic review, described earlier in Chapter 2, provided evidence that a vast majority of approaches to SA in health and well-being is based on rule-based and traditional machine learning approaches, whose results are sub-par to those achieved in other domains. These approaches require features to be engineered manually. In contrast, deep learning applies layers of linear and non-linear data transformations to learn a representation of the problem that is best suited for the end task. Raw natural language text is usually represented as a sequence. Therefore, the use of RNN to process such text comes as a logical choice, as this architecture is specifically designed to process sequences. Not surprisingly, RNNs have been employed to perform both general SA and aspect-based SA. As stated previously, simple RNNs suffer from the vanishing gradient problem, which LSTMs overcome.

Liu et al. [163] utilised an LSTM to perform SA of movie reviews. They proposed three different ways to jointly train the network on multiple tasks. First, multiple tasks shared the same LSTM layer. Second, each task was assigned a single LSTM layer where these layers shared the information. Third, each task was assigned a separate LSTM layer but a BiLSTM layer was used to capture the shared information. The output of these layers is later fed to the output layer, which differs depending on a given task. In [209], a backward LSTM (from right to left) was used to encode negators and intensifiers, after which it was passed to a BiLSTM for sentiment classification.

Different deep learning architectures, although most commonly LSTM, have been applied to solve the problem of aspect-based SA. A groundbreaking approach employed hierarchical BiLSTM on constituency parse trees to perform aspect-based SA on restaurant and laptop reviews [220]. Similarly, an LSTM approach achieved the accuracy of 84% and 90% on classifying customer reviews into positive and negative sentiment, respectively [251]. In this work, the embedding of the aspect was concatenated to the embedding of each individual word in order to focus attention to the aspect. In that way, the model focused on those parts of the text that were affected by the aspect. More recently, lexicons were used to obtain more lexical features, which were combined with the attention that was learnt from the output of the LSTM to perform aspect-based SA of restaurant reviews in [54], achieving the accuracy of approximately 83%. In [260], two approaches were proposed to exploit sentiment lexicons for SA. First, lexicons were used to learn sentiment-aware attentions to highlight important words. Second, the model was used to learn sentiment word embeddings. Both outputs were later passed through the BiLSTM for sentiment classification. In [116], the authors used a bidirectional GRU, similar to LSTM yet simplified in their complexity, to perform aspect-based SA of drug reviews. Pre-trained weights learnt from short texts were used to initialise the weights of two BiGRUs, which were then trained to learn the representation of the text and that of the aspect. However, the accuracy of this approach was below 80% on drug reviews [116]. Although such underperformance may be attributed to differences in architectures used and specific properties of the training data, it is in line with a previous finding from Chapter 2 that SA in health and well-being does lag behind the state of the art in other domains. This is mainly due to the generally negative connotation of health-related concepts, which tends to skew the results of SA toward negative polarity as described in Chapter 3. It is, therefore, ever so important to carefully examine the context when such concepts are used as aspects in SA.

An attention mechanism can be used to improve the performance of RNNs in aspect-based SA by letting them know where to focus their learning. An attention-based bidirectional CNN-RNN provides a hybrid model in which bidirectional RNNs are used to model both long and short contextual dependencies, local features robust to positional changes are selected using CNNs and an appropriate emphasis is placed on different words by applying the attention mechanism on the output of bidirectional layers [55].

RNNs are optimised to process sequences and, therefore, are not ideally suited for context-sensitive tasks such as aspect-based SA. CNNs are better suited to represent contextual information and as such have been used in SA applications as local feature extractors [274, 52]. In [146], it was demonstrated that CNNs can achieve great results on multiple downstream tasks including the binary sentiment classification of movie reviews. The model architecture is shown in Figure 4.5. The model contains multiple convolutional filters with different widths that are applied directly to the word embeddings. This is similar to the way convolutional filters are applied in computer vision except in one dimension instead of two. These filters produced different feature maps that were subsequently used for sentiment classification. However, one downside of CNNs in NLP is their inability to model long distance dependencies. This architecture is instead more suitable for representing the close context of the words mainly resulting in modelling n-grams.

In [250] authors proposed a hybrid architecture that combined CNN and LSTM to classify the sentiment of short text by learning local features with CNN and long distance dependencies with LSTM. First, the convolution was performed on a sentence level, after which the output of the CNN was passed through an LSTM. Similarly, another study [181] used an ensemble of CNN and BiLSTM. The output of both layers was averaged in order to classify the sentiment of a document.

A CNN was used in combination with lexicons in [227]. They were in fact integrated in three different ways including naive concatenation, multichannel integration and separate convolution. Naive concatenation represents the concatenation of lexicon embeddings to the word embeddings. Multichannel integration uses two channels, one to represent word embeddings and the other to represent lexicon embeddings. Separate convolution represents two individual convolutions applied to word and lexicon embeddings, respectively. After the convolution, they are concatenated and passed through a softmax layer.

The previous approaches may struggle when encountering words that have not been seen during training. However, CNNs do not necessarily need to operate on individual words. Instead, they can be applied at a character level [280]. Such an approach overcomes the problem of out-of-vocabulary words. The characters are represented with a one-hot encoding vector, after which the sequences of characters are passed to the CNN to classify the sentiment.

Both RNN and CNN approaches, mentioned above, operate on text represented as a sequence. Such representation is not ideally suited for aspect-based SA. Namely, the sentiment of a specific aspect is directly influenced by its modifiers and not necessarily the entire context, which may end up adding unnecessary noise to the problem representation. Proximity and order are often used as proxies in lieu of explicit dependencies. Therefore, most often, context will be represented using n-grams or sequences. Constituency parse trees also take advantage of the notions of proximity and order to group words together into coherent phrases, which represent key features for semantic analysis and thus can support applications such as question answering and information extraction [140]. However, applications such as aspect-based SA depend crucially on the use of modifiers, which can change or emphasise a particular word in a sentence. Universal dependencies represent grammatical relations between words in a sentence [86] in the form of triplets (name of the relation, governor and dependent), which give rise to a graph representation of a sentence, which is often collapsed into a tree [87]. In aspect-based SA, such structure can be used to explore those words that are logically associated with the given aspect regardless of their physical proximity and ignore (or downplay) those that are less relevant with respect to the expression of sentiment.

The aspect's relations to other words represent important features of its sentiment, but are not taken into account in RNN-based approaches. Although raw text is represented as a sequence of either characters or words, with additional linguistic processing, its representation can be converted into a graph. Unlike CNNs which are specialised to perform convolution over a grid structure, graph convolution enables the model to perform convolution over arbitrary graphs. GCN, which performs convolution over a graph, is a most naturally suited neural network architecture for this type of sentence representation. GCN acts as a message passing algorithm, where the information between vertices in the graph is propagated along the edges, allowing the vertices to aggregate information from their neighbours. In NLP, this type of architecture has previously been employed to extract semantic relations from syntactic dependency parse trees in [281]. Most recently, GCN has been used to perform aspect-based SA on product reviews in [238] and on Twitter data in [53]. In [53] GCN was additionally combined with the syntactic dependencies.

A GCN approach may not be able to capture the features of long-distance dependence, thus struggling to effectively represent the aspect's context. This issue can be easily resolved by adding transitive edges to the dependency graph, which has been proven to improve the representation of sentiment dependencies [283]. Alternatively, a phrase dependency graph can be constructed by integrating the constituency and dependency parse trees [261]. Further embellishing the dependency graph by leveraging information from a sentiment lexicon was found to improve the learning ability of a GCN model in aspect-based SA [159]. However, adding more information may introduce noise and inefficient use of information relevant to SA. Namely, despite the direct or indirect connection with an aspect in the dependency tree, only few words add value to predicting the sentiment polarity of the aspects. These words tend to be adjectives and verbs. Therefore, part-of-speech information can be used to prune the dependency tree with two benefits [264]. First, fewer unrelated words are connected directly or indirectly to the aspect, which reduces the noise they bring to the convolution. Second, a more concise syntactic dependency graph leads to fewer convolutions, thus making the corresponding GCN more efficient.

SA suffers from domain dependency [63]. On one hand, it requires a lot of training data. In particular, deep learning algorithms are known to be data hungry. On the other hand, when a SA model trained on one domain is applied to a different one without any transfer of knowledge, the performance tends to deteriorate. One way to tackle the problem of domain shift is to create an ensemble of models trained on different data sources [198]. An ensemble of models whose individual predictions are combined in a way in which the given models compensate for each other's weaknesses [144]. In particular, heterogeneous ensembles use different learning algorithms to generate different types of base classifiers. Recent experiments in SA demonstrated that ensemble learning can improve the accuracy. For example, the stacking of LSTM, CNN and CNN-BiLSTM and support vector machine (SVM) significantly improved the accuracy of SA in Chinese albeit failing to replicate the success in English [171]. Nonetheless, another study, which widened the choice of base classifiers to four pre-trained lexicon-based models and six machine learning algorithms (naïve Bayes, SVM, logistic regression, feedforward neural network, CNN and LSTM), managed to improve performance by more than 5 percent points over the best individual model [144]. The true potential of ensemble approaches to SA lies in leveraging symbolic models (such as lexicons and grammatical relationships) to encode meaning and subsymbolic methods (such as word embeddings and neural networks) to infer patterns from data [63].

More recently, the research attention has shifted towards large pre-trained language models such as BERT [92]. BERT is a transformer-based architecture that provides contextual word embeddings; it uses an attention-based mechanism, rather than recurrence, to determine which words are important for the overall context within the document [245]. It has enabled great performance improvements across a variety of NLP tasks. The main advantage of BERT is that it can easily be fine-tuned using additional training data to solve specific NLP tasks such as aspect-based SA. The advantage of large pre-trained language models such as BERT is that it requires less data for fine-tuning. This is of particular importance in the domain of health and well-being, where privacy concerns limit data availability and in turn wide-spread adoption of deep learning [231]. Moreover, the annotation of such data often requires the assistance of medical professionals in contrast with annotation in many other domains, where annotation can be crowdsourced from non-experts.

In addition, BERT is commonly pre-trained for specific domains to improve its performance on different sublanguages. For example, of relevance to the domain of health and well-being are BioBERT [154] and ClinicalBERT [46]. However, when lay language is processed in this domain, BERT's performance may still be superior to the special-trained language models. For example, when BERT was used to understand people's opinion towards vaccination, multilingual BERT model outperformed both BioBERT and ClinicalBERT [50]. BERT was successfully fine-tuned to perform SA of drug reviews [192], albeit without focusing on specific aspects.

When fine-tuning BERT, the task of aspect-based SA can be formulated as a question answering task, where the aspect represents a question and its sentiment is the answer. BERT typically represents this task by pairing up two sequences, one representing the source sentence and the other one specifying the phrase that corresponds to the aspect. This approach was successfully adapted to classify the sentiment associated with a specific aspect of a product or a restaurant expressed in their reviews [237, 267, 122, 157]. Such an approach improved the results relative to the models that use a single sequence to perform aspect-based SA [237].

The application of deep learning to the problem of SA in the domain of health and well-being has not been studied as extensively as to the problem of SA in other domains. To date, SA in health and well-being has been mostly focused around drug reviews due to the data availability [151, 116, 80]. In this thesis a new application of neural networks for the task of aspect-based SA in health and well-being is proposed. The following two Chapters provide the details of two deep learning approaches we developed to support aspect-based SA in health and well-being.

Chapter 5

A Graph Convolutional Approach to Aspect-Based Sentiment Analysis

The work presented in this Chapter has been published in the journal of Artificial Intelligence in Medicine. It is based on the following paper: Žunić Anastazia, Corcoran Padraig and Spasić Irena. Aspect-based sentiment analysis with graph convolution over syntactic dependencies. *Artificial Intelligence in Medicine*, 119, 102138, 2021.

5.1 Introduction

As previously stated in the thesis, sentiment analysis (SA) is an NLP task which aims to classify the sentiment expressed or implied by a given piece of text. It can be applied at different levels of text organisation: the whole document [269], an individual paragraph or sentence, or a specific aspect [116]. In this Chapter, we specifically focus on aspect-based SA. This task is particularly difficult in the domain of health and well-being, where the performance of SA was found to lag behind the state of the art, which has previously been described in Chapter 2.

Recent proliferation of online platforms designed to share health-related information with other users sparked research interest in SA in this domain. However, existing research in aspect-based SA is typically conducted using user reviews of products and services such as mobile devices and restaurants, but also pharmaceutical drugs, which are related to one's health and well-being.

In the case of drug reviews, current research efforts in SA are focusing on the whole document (i.e. review) and not on an individual aspect. This is partly related to the availability of annotations that can be used to train supervised classification approaches. Reviews typically come together with star rating, which can be easily converted into sentiment labels. Aspect-based SA requires manual data annotation, which has been identified as one of the key obstacles to machine learning approaches in clinical NLP [231]. SentiDrugs [116] represents a dataset relevant to the current project: it consists of drug reviews in which aspects were manually identified and annotated for sentiment. Unfortunately, this dataset is not publicly available.

In terms of methods used to support SA in health and well-being, our systematic review revealed that rule-based and traditional machine learning approaches are used most commonly (see Chapter 2). Both approaches require manual engineering of either rules or features, which limit their portability across different tasks and domains. On the other hand, deep learning does not suffer from these limitations and has demonstrated considerable success in a variety of NLP tasks including SA. For example, deep learning has been successfully applied to support SA of user reviews of hotels and restaurants as well as product reviews (phones, cameras, laptops, etc.) [220, 251]. Incorporating syntactic structure into the deep learning process to model sentiment compositionality can improve the performance of SA by almost 10 percent points as confirmed by Socher et al. [228] who trained a recursive neural network on constituency parse trees. Dependency parse trees of sentences directly capture syntactic dependencies between the words and in that respect may be better placed to support aspect-based SA by traversing dependencies associated with the target word or phrase. From the existing variety of neural network architectures, graph convolutional networks (GCN) are most naturally suited to traversing the graph structure of syntactic dependencies.

In this Chapter, we hypothesise that a GCN approach should outperform traditional neural network architectures on the task of aspect-based SA. To test this hypothesis, we developed a new approach to aspect-based SA based on graph convolution, where the aspect of interest is represented as a vertex within the graph representation of a sentence, and its representation is a result of convolutions performed along its edges and those of its neighbours. Additionally, we create a new dataset to evaluate this approach. To examine the effect of syntactic dependencies on sentiment polarity of a given aspect, we tested the proposed approach on two graph representations of a sentence including a simple sequence and a dependency parse tree. The experiments asserted the importance of features incorporating syntactic dependencies over sequential order for aspect-based SA.

This Chapter is organised as follows. Section 5.2 describes the methodology. Data collection and the evaluation of the approach are described in Section 5.3. Finally, Section 5.4 summarises our findings.

5.2 Methodology

Aspect-based SA is a fine-grained SA task, where the goal is to identify the sentiment of the specific aspect, rather than the overall sentiment of the document. In this Chapter we focus on binary classification of the given aspect, where the goal is to classify the sentiment of the given aspect into one of the two classes, positive or negative.

The overall system design is shown in Figure 5.1. A set of drug reviews, originally published in [109], was collected from Drugs.com [10], which is the largest, most widely visited, independent pharmacologic drug information website, that allows its users to post reviews describing personal experience about their use. All reviews were automatically annotated with concepts from the Unified Medical Language System (UMLS), a large repository of inter-related biomedical concepts and the corresponding terminology [58]. These concepts represent aspects whose sentiment needs to be classified. Input documents are processed by Stanford CoreNLP [68] to convert individual sentences into dependency parse trees, i.e. dependency graphs. Individual

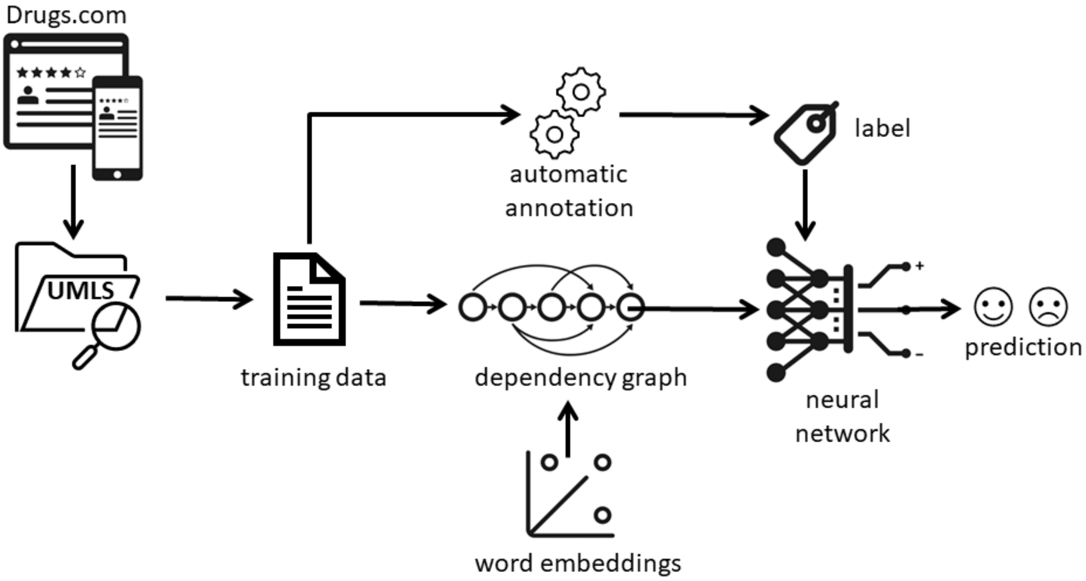

Figure 5.1: System design for aspect-based SA of drug reviews.

words representing vertices in such graphs were mapped onto their embeddings, which were pretrained on web data from Common Crawl using the *GloVe* model [205]. Each input sentence is represented as a sequence of tokens $S = (w_1, w_2, ..., w_n)$. When combined with word embeddings, this representation gives rise to a matrix of dimensions n x d, where n is the total number of tokens and d is the size of the embedding vector space. These data were combined with sentiment labels to train a neural network to classify the sentiment associated with individual UMLS concepts. The architecture of this neural network, shown in Figure 5.2, was based on graph convolution. The input consists of the dependency graph of the sentence whose vertices are convoluted by propagating information from other vertices across the edges of the graph. After two successive convolutions, the vertex corresponding to the aspect is mapped onto 2-dimensional classification space whose dimensions correspond to positive and negative polarity, respectively. This means that the two polarities produced as an output are extracted directly from the aspect and indirectly from its neighbours via convolution, and therefore their values may vary across different aspects within the same sentence.

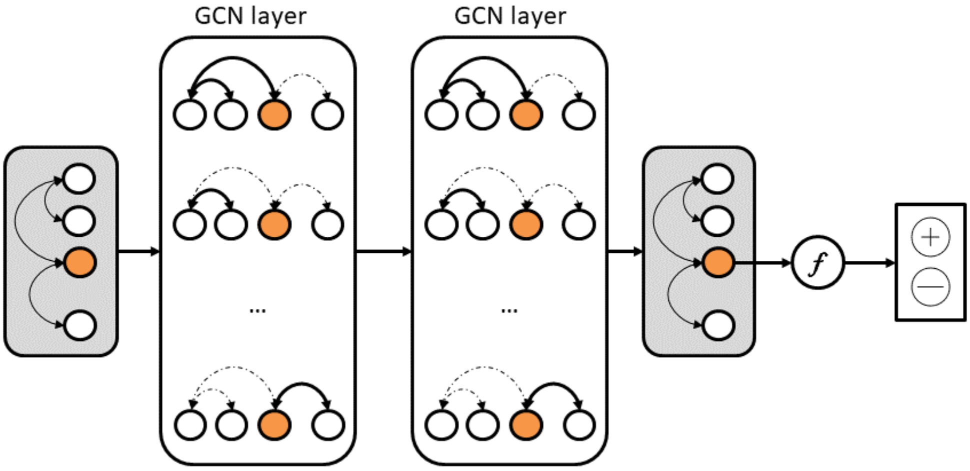

Figure 5.2: Graph convolutional network over dependency graphs. The sentiment of an aspect, highlighted in the graph, is classified.

5.2.1 Problem representation

Sentence S is a sequence of tokens that can be represented as a graph $G = (V, E)$. Graph is an ordered pair (V, E), where V is a set of vertices (nodes) and E is a set of edges that represent pairs of vertices. Graph $G = (V, E)$ of a sentence can be modelled in different ways, for example sequence graph, dependency graph, relational dependency graph, or constituency graph. Vertices represent the tokens within the sentence, and edges connect the vertices, i.e. tokens. The structure of sequence, dependency and relational dependency graph are shown in Figure 5.3.

We decided to represent a sentence using its dependency parse tree, which is a special case of a directed dependency graph. Here, each vertex has a single parent ex-

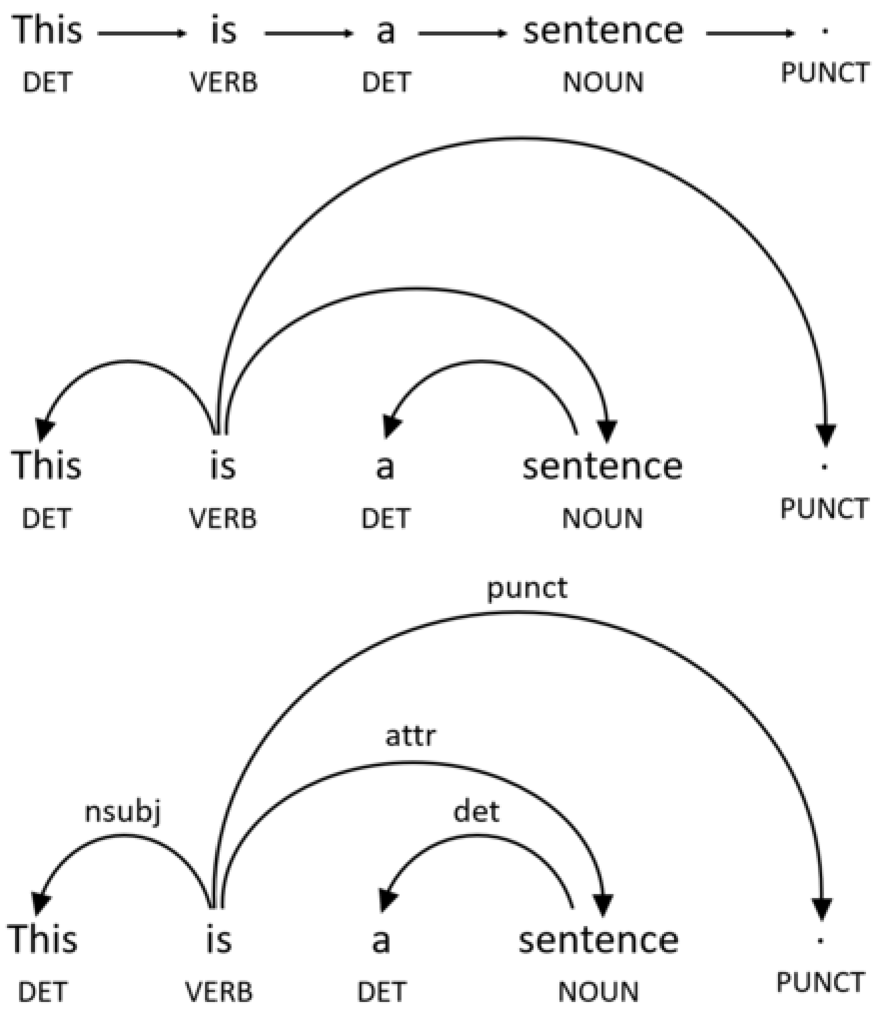

Figure 5.3: Three types of graphs that are considered for sentence representation: sequence graph, dependency graph, and relational dependency graph with different edge types from top to bottom, respectively.

cept for the root, which does not have a parent. Figures 5.4 and 5.5 provide illustrations of dependency parse trees produced automatically by Stanford CoreNLP [68] and spaCy [125], respectively. These software packages employ different algorithms for sentence analysis. Therefore, they do not always provide the same output as can be seen from Figure 5.5. In this particular case, the dependency parse tree produced by Stanford CoreNLP is not entirely correct. This is an important point to note. Some incorrect parses may actually have better utility for the purpose of aspect-based SA.

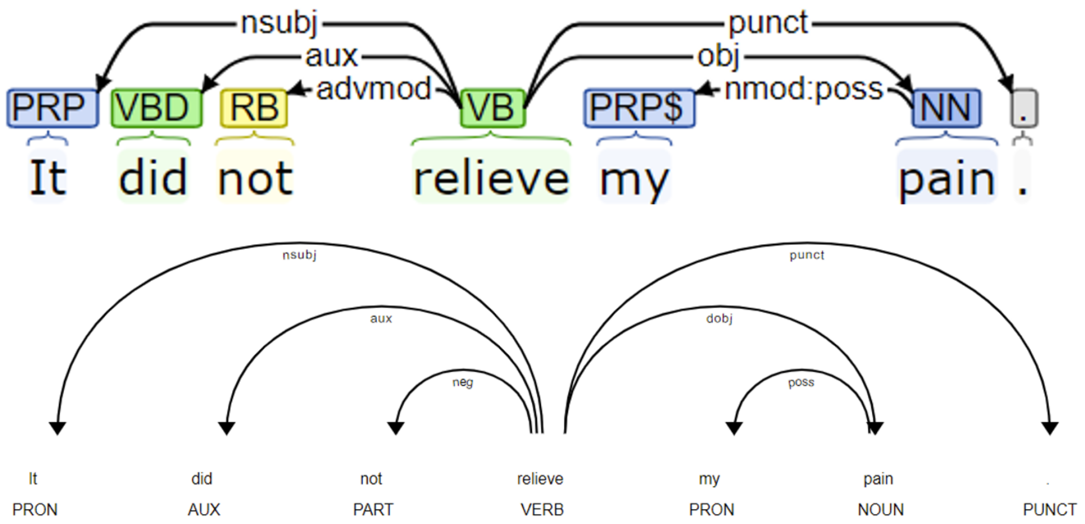

Figure 5.4: Example of negation in a dependency parse tree produced by Stanford CoreNLP and spaCy, respectively.

For example, the incorrect parse in this case places the word "relief" closer to the aspect, one degree of separation in Stanford CoreNLP versus two degrees of separation in spaCy. In any case, the GCN should adapt to such noise in the training data to maximise the utility for SA. Lastly, in order for the convolution to flow in both directions across a dependency graph, we chose to ignore the direction.

5.2.2 Graph convolution

Graph convolutional network (GCN) is a neural network architecture that takes a graph as an input and outputs an updated representation of each vertex in the given graph. The updated representation of each vertex is a function of the current representation of the given vertex and its neighbouring vertices. Various GCN architectures have been proposed [263]. In this Chapter, we used two types of GCN architectures, which were applied to the sentence dependency parse tree. The architectures in question are *GraphSAGE* GCN (GS-GCN) [115] and relational GCN (R-GCN) [225]. These are commonly used GCN architectures which have been found to provide good perform-

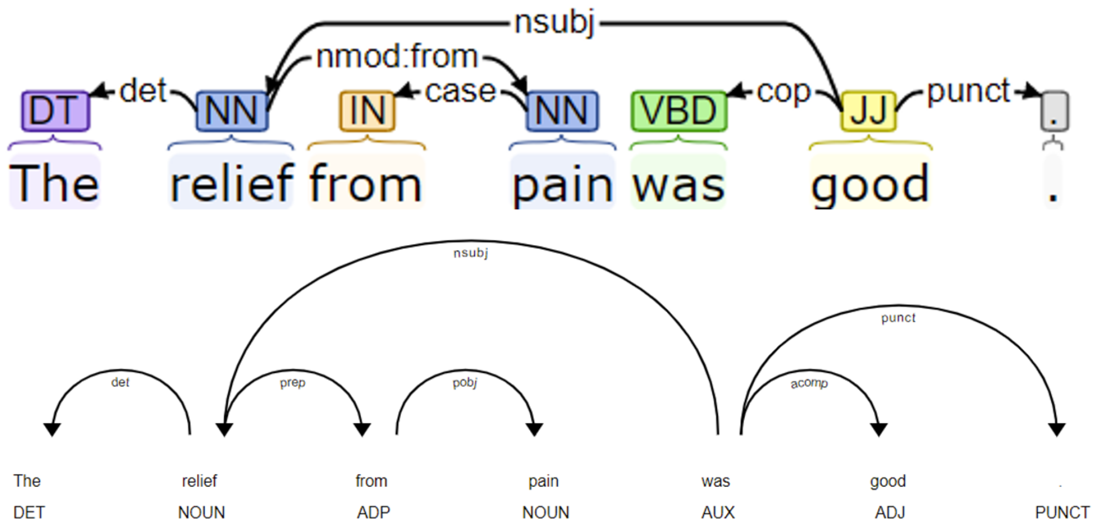

Figure 5.5: Example of the dependency parse tree produced by Stanford CoreNLP and spaCy, respectively.

ance in empirical comparisons of different GCN architectures [102]. We now describe the details of each GCN architecture considered in this Chapter.

The input to each GCN architecture is the undirected dependency graph of the sentence, where each vertex is initialised with the corresponding 300-dimensional word embedding vector. Vertex representation is updated with each convolution over the graph. A sequence of k convolutions will propagate information across the graph to the k-th order neighbour. Our architectures each contain two graph convolution layers, which means that every vertex will get information from the second order neighbour. Specifically, the first convolution layer reduces the dimensionality of the input vectors from 300 to 125 and the second layer will output a 100-dimensional vector. In the GS-GCN architecture the hidden state h_i^t corresponding to vertex i and layer t after one convolution is updated as follows:

$$h_i^t = \sigma(W^t \cdot concat(h_i^{t-1}, aggregate(h_j^{t-1}, \forall j \in N(i)) + b^t) \tag{5.1}$$

where $N(i)$ denotes a set of neighbours of vertex i, $\sigma(\cdot)$ is a nonlinear function,

$concat(\cdot)$ represents concatenation of vectors horizontally and $aggregate(\cdot)$ indicates the summation of the neighbours of the corresponding vertex. Moreover, W^t and b^t are weight matrix and bias, respectively, for the t-th convolutional layer, which are the parameters of the model that are learned. Non-linear function $\sigma(\cdot)$ in this architecture is a rectified linear unit (ReLU). The potential downside of this method can be that the edge types are not taken into consideration. The solution to that is to use another graph convolution algorithm known as relational graph convolution (R-GCN) [225]. In order to perform R-GCN, a graph is represented as a triplet $G = (V, E, R)$, where R is a set of relations r_i corresponding to each edge e_i. In the R-GCN architecture the hidden state h_i^t corresponding to vertex i and layer t after one convolution is updated as follows:

$$h_i^t = \sigma \left(\sum_{r \in R} \sum_{j \in N_i^r} \frac{1}{c_{i,r}} W_r^{t-1} h_j^{t-1} + W_0^{t-1} h_i^{t-1} \right) \tag{5.2}$$

where N_i^r denotes a set of neighbours of vertex i that are connected with relation $r \in R$, $c_{i,r}$ is a normalisation constant. Relations r are directly extracted from the output of the Stanford CoreNLP dependency parser. Figures 5.4 and 5.5 provide examples of different edge types including *nsubj* - nominal subject, *neg* - negation modifier, *obj* - object, etc. More detailed explanation about the syntactic relations can be found in Universal Dependencies documentation [87, 33]. An overview of the different graph types used in our experiments for sentence representation is shown in Figure 5.3.

5.2.3 Classification

After two successive graph convolutions, the hidden state of the sentiment aspect is retrieved and mapped into the classification space by a linear transformation, reducing the dimensionality of its embedding from 100 to 2, where the two dimensions correspond to positive and negative polarity, respectively. The 2-dimensional vector is then passed through the softmax layer, which provides the probability distribution over the two sentiment polarities.

5.3 Results

To validate the model proposed in this Chapter we performed two sets of experiments. The first set compares different types of graphs as input to the GCN models, whereas the second set compares different neural network architectures including GCN, RNN and LSTM. All methods were implemented in the Python programming language using the PyTorch library for deep learning. The source code is available at `https://github.com/zanastazia/ABSA-with-GC-over-Syntactic-Dependencies`.

5.3.1 Data

To train the model and evaluate its performance, we created a dataset specifically for the task of aspect-based SA. It consists of drug reviews borrowed from another study [109], which were publicly available from the UCI Machine Learning Repository. A total of 128,581 reviews were originally collected from the Drugs.com website [10]. Each review comes with a star rating on a scale from 1 to 10, which was converted into a sentiment label. This dataset was previously annotated for aspect-based SA [116], but unfortunately it is not publicly available.

The choice of aspects in this Chapter was motivated by the likely practical applications of SA on this dataset. The most obvious applications are related to the drugs' efficacy and safety, which could be inferred from the sentiment associated with the signs and symptoms discussed in drug reviews. Therefore, we chose signs and symptoms as aspects of SA. Aspects were annotated automatically using relevant concepts from the Unified Medical Language System (UMLS), a large repository of inter-related biomedical concepts and the corresponding terminology [58]. We focused on a single subclass of UMLS concepts that represents clinical signs and symptoms [242]. We then cross-

lexicon	description
AFINN [187, 2]	A list of 2477 words and phrases with an integer value as a sentiment score between -5 (negative) and 5 (positive).
EmoLex [184, 22]	A lexicon where items are annotated with 8 basic emotions and sentiment score of either positive or negative.
Harvard General Inquirer [235, 13]	Lexicon that provides 1915 positive and 2219 negative words.
MPQA [257, 18]	A lexicon of around 8000 items that provides sentiment scores.
Opinion lexicon [131, 23]	A list of approximately 6800 positive and negative words.
WordNet Affect [236, 36]	An extension of WordNet, each item in the lexicon is labelled as positive, negative, ambiguous, or neutral.

Table 5.1: A selection of sentiment lexicons.

referenced these concepts against six lexicons described in Table 5.1 to identify those with negative sentiment. This choice was based on a previous finding that the negative connotation of health symptoms tends to skew the SA results toward the negative spectrum (see Chapter 3). In other words, the sentiment of such aspects is more challenging to classify. In this study, we want to focus specifically on this bias by exploring the ways in which the context (represented by syntactic dependencies) can modify the negative polarity associated with signs and symptoms.

The most frequently mentioned signs and symptoms were selected to represent aspects whose sentiment needs to be classified: *burning, constipation, dizziness, dizzy, dry, fatigue, headache, nausea, nauseous, nauseated, pain, painful, sick, sickness, symptom, tired* and *tiredness*. The silver-standard sentiment of each aspect was inferred from the corresponding user review's star rating ranging from 1 and 10. To easily convert star rating into sentiment, we annotated reviews with star rating of 1 or 2 with negative sentiment, those with rating of 9 or 10 with positive sentiment and removed the remaining reviews from further consideration. The overall distribution of sentiment before the selection of short reviews is shown in Figure 5.6.

The sentiment may vary across a long document. To increase the likelihood of the overall sentiment being related to a specific aspect, only short reviews, specifically

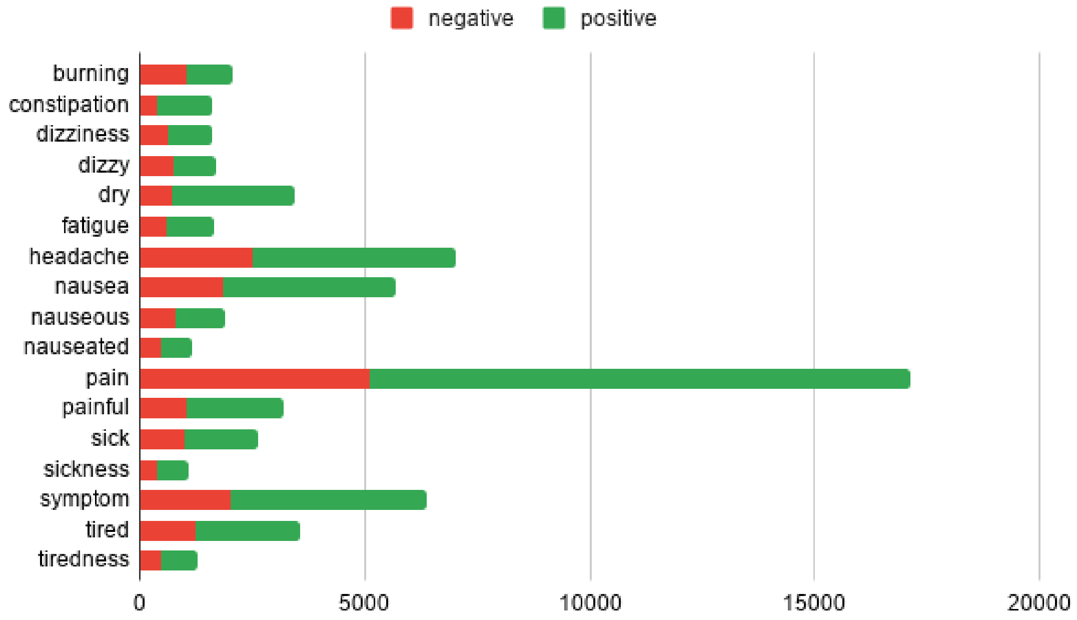

Figure 5.6: Distribution of aspect-based sentiments.

those consisting of a single sentence, were considered. This reduced the need for manual annotation albeit at the cost of reducing the size of the dataset. As a result, we ended up with a set of 806 positive and 612 negative reviews. These reviews were curated manually by reading them and checking whether they were correctly annotated. Incorrectly annotated reviews were removed. As a result, 79.28% of positive reviews and 96.89% of negative reviews were retained. The choice between removing and re-annotating the incorrectly annotated reviews was motivated by the need to balance the dataset. Ultimately, a total of $1,232$ reviews were retained out of which 639 (51.87%) were positive and 593 (48.13%) negative, which represents a well-balanced set despite the fact that the chosen signs and symptoms are otherwise inherently negative. Finally, each aspect (i.e. a reference to a sign or symptom from Figure 5.6) was mapped to the sentiment of the corresponding sentence. Table 5.2 provides a sample of annotated sentences with aspects indicated by bold typeface.

Data were split randomly to use 80% for training and keep the remaining 20% for testing. From the training subset, 20% was used to tune hyperparameters prior to training

Sentence	**Sentiment**
Excellent **headache** reliever!	$+$
Good medicine, it gets rid of your **pain** without that drowsy sick feeling.	$+$
Love this medicine, no **headache**.	$+$
Sadly no effect on my **pain**.	$-$
Made my **symptom** worse - so much for 24 hour relief.	$-$
No **pain** relief whatsoever.	$-$

Table 5.2: A sample of sentences with their aspects (in bold typeface) annotated with the corresponding sentiment.

the model on the remaining 80%. The distribution of sentiment within training, validation and testing is provided in Table 5.3. No substantial differences in the distribution of the two sentiment labels across the training, validation and test sets can be observed.

	positive	negative	total
train	410	378	788
validation	99	98	197
test	130	117	247
total	639	593	1232

Table 5.3: The distribution of sentiment in the training, validation and test sets.

5.3.2 Hyperparameters

The model was trained with backpropagation by minimising cross-entropy loss function, and optimised with the Adam optimizer [147]. Learning rate was set to 0.001. To avoid overfitting we applied early stopping, where the training was stopped if the validation loss increased in five consecutive epochs, therefore the patience was set to

5. Apart from early stopping, dropout was used as regularisation and set to 0.2, which means that during the training process 20% of the weights were set to zero.

5.3.3 Evaluation measures

The measures used to evaluate the performance of the model included accuracy, F-measure and cross-entropy loss. As mentioned in the previous Chapter, the evaluation measures are defined as follows. Accuracy is calculated as the percentage of correctly classified instances. Precision (P) and recall (R) are calculated using true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) using the equations from the Chapter 4, see Equation 4.2. These two values are combined into the F-measure as follows: $F - measure = 2 * P * R/(P + R)$. Cross-entropy loss for binary classification is defined as:

$$loss = -\frac{1}{n}\sum_{i=1}^{n} ln(p_i) \tag{5.3}$$

5.3.4 Sentence representation

To investigate whether a dependency graph is better suited for modelling aspect-based SA than a simple sequence (also represented by a graph, see Figure 5.3), we used both and compared the results. In both cases, we performed experiments with both directed and undirected graphs. The results achieved using the GS-GCN model over different types of input graphs are reported in Table 5.4. The best performance across all considered measures were achieved when an undirected dependency graph was used for sentence representation.

graph type	loss	F-measure	accuracy (%)
sequence	0.6348	0.6529	65.59
undirected sequence	0.5635	0.7332	73.68
dependency	0.5953	0.7001	70.04
undirected dependency	**0.4570**	**0.8179**	**81.78**

Table 5.4: Evaluation of the GS-GCN architectures over sequence and dependency graphs with different edges.

5.3.5 Neural network architectures

To establish the baseline, we also performed experiments with two architectures, RNN and LSTM, which have previously been used to support aspect-based SA [251, 54]. They used a sequence, which is equivalent to a sequence graph we defined earlier. Note that, CNN was not considered because it operates on a feature vector used to represent the whole sentence rather than each token individually [146].

Apart from the standard RNN, LSTM and two GCN architectures (GS-GCN and R-GCN), we also performed experiments with BiLSTM-GS-GCN, in which the input was passed through a BiLSTM as proposed in [238] prior to performing graph convolution. These models were trained using the Adam optimizer with a learning rate of 0.001. Dropout of 0.2 and early stopping with patience of 5 were applied to reduce overfitting. The results are presented in Table 5.5. The four architectures based on graph convolution outperform the rest. The best results were achieved using the GS-GCN over the undirected dependency graph. These results indicate that a dependency graph enriches sentence representation, which in turn enables the model to exploit syntactic dependencies for semantic reasoning.

We used the Z-test for the equality of two proportions [143] to test statistical significance of differences in accuracy between GS-GCN against that of the five baseline methods, respectively. The null hypothesis is that there is no significant difference between the architectures in terms of the accuracy they have achieved. Table 5.6 shows that all p-values are ≤ 0.05, hence we reject the null hypothesis and conclude that there is significant difference between the given architectures in terms of their accuracy. Therefore, the GS-GCN over the undirected dependency graph is indeed significantly more accurate than any of its counterparts.

Evaluation of the baseline architectures and of the proposed model over positive and negative sentiments can be seen in Table 5.5. Most of the architectures perform better on classification of positive sentiment. This trend was overturned by GCN over the dependency graph, while still providing the best results on classification of positive sentiment, suggesting that this approach makes the best utilisation of context to perform aspect-based SA.

5.4 Summary

We proposed a new approach to aspect-based SA based on graph convolution over the dependency parse tree of the sentence. The experimental results show that, relative to other neural network architectures and sentence representations, this approach makes the best utilisation of context to perform aspect-based SA. We specifically looked at the sentiment surrounding medical signs and symptoms because of the negative sentiment underlining their semantics, which makes SA in the domain of health and well-being challenging. As mentioned before, for someone suffering from a chronic condition, having a good quality of life is not necessarily measured by the absence of associated signs and symptoms, but rather by the extent to which they can be successfully managed and controlled. However, the negative connotation of signs and symptoms tends to skew the results of SA toward negative polarity, described earlier in Chapter 3. This is one of the reasons SA in health and well-being is performing below the F-measure of 60% on average, lagging behind the state of the art in SA on service and product reviews, where F-measure is found to be above 70% and 80%, respectively (see Chapter 2). In this Chapter, we successfully tackled this bias by exploring the ways

method	**sentence representation**	**loss**	**F-measure**	**acc (%)**	**pos acc (%)**	**neg acc (%)**
RNN	sequence	0.6202	0.6725	67.61	76.92	57.26
LSTM	sequence	0.6259	0.6725	67.61	76.92	57.26
R-GCN	undirected dependency	0.6265	0.6833	68.42	76.92	58.97
GS-GCN	undirected sequence	0.5635	0.7332	73.68	77.85	68.39
BiLSTM+GS-GCN	undirected dependency	0.5095	0.7566	75.71	76.92	70.09
GS-GCN	undirected dependency	**0.4570**	**0.8179**	**81.78**	**78.46**	**85.47**

Table 5.5: Evaluation of the baseline models and of the proposed model (GS-GCN-undirected dependency).

method	accuracy (%)	p-value
RNN	67.61	0.0001
LSTM	67.61	0.0001
R-GCN	68.42	0.0003
GS-GCN (undirected sequence)	73.68	0.0152
BiLSTM+GS-GCN	75.71	0.0496

Table 5.6: Comparison of the GS-GCN over the undirected dependency graph with accuracy of 81.78% **against the baseline methods.**

in which the context (represented by syntactic dependencies) can modify the negative polarity associated with signs and symptoms and effectively closed this performance gap by achieving state-of-the-art results regardless of the domain.

Chapter 6

A Transformer-Based Approach to Aspect-Based Sentiment Analysis

The work presented in this Chapter has been published in the journal of Artificial Intelligence in Medicine. It is based on the following paper: Žunić Anastazia, Corcoran Padraig and Spasić Irena. The case of aspect in sentiment analysis: Seeking attention or co-dependency? *Machine Learning and Knowledge Extraction*, 4(2), 474-487, 2022.

6.1 Introduction

In Chapter 5, we described a novel neural network architecture based on graph convolution which is applied directly to a dependency parse tree. This approach improved the performance of aspect-based SA in health and well-being relative to other standard neural network architectures. These results suggest that the features incorporated within the dependency parse tree encode important information for the classification of sentiment. In the meantime, a new neural network architecture called a transformer [245] started dominating the field of NLP. This architecture has been used to successfully train very large language models that produce contextualised word embeddings. In our previous approach we used graph convolution to contextualise the aspect of SA. Naturally, a research question emerged around comparing the two ways of contextualising the aspect and their effects on classifying its sentiment. Therefore, we developed an alternative approach to aspect-based SA based on a transformer architecture.

The most popular architecture of this kind is called Bidirectional Encoder Representations from Transformers (BERT) [92]. Its popularity lies in the fact that it can not only be pre-trained to generate contextualised word embeddings but can also be easily fine-tuned using relatively small datasets to support downstream NLP tasks such as that of SA. Therefore, in this Chapter we investigate the potential of a BERT-based approach to aspect-based SA in the domain of health and well-being.

This Chapter is organised as follows. Section 6.2 describes the methodology including implementation details and model training. Evaluation of the model and its comparison to the baseline established in Chapter 5 is given in Section 6.3. Section 6.4 discusses possible interpretations of the results. Finally, Section 6.5 summarises our findings.

6.2 Methodology

The goal of aspect-based SA is to classify the sentiment of a document with respect to a particular aspect. Therefore, the document and the aspect considered constitute the input, whereas the output represents the sentiment classified into one of two classes, positive or negative. The first step in fine-tuning BERT for this particular task is choosing an appropriate representation of the problem.

6.2.1 Neural network architecture

BERT [92] is a transformer-based language model, which utilises transformer encoders to create contextualised word embeddings. Transformers [245] are based on an encoder-decoder neural network architecture that uses attention mechanism. Note that, transformers process the data simultaneously, rather than sequentially, as it is the case with recurrent neural network architectures such as LSTM. The self-attention layer considers all words, each represented by its embedding and its position relative to other words, to improve its encoding of the entire sentence. In other words, self-attention determines the impact of individual words on the sentence interpretation.

BERT is pre-trained on two tasks simultaneously, masked language modelling and next sentence prediction. When performing masked language modelling, BERT hides a certain percentage of words by using a special token [MASK] instead and uses their position to infer these words. By performing this task BERT learns relationships between words. Whereas, when performing next sentence prediction, BERT learns long-term dependencies across sentences. BERT uses two special tokens to support fine-tuning and specific task training. The first one is a classification token [CLS]. It indicates the beginning of a segment, typically a sentence, and is commonly used for classification tasks, hence the name. The output associated with this token is used to make a prediction about the given segment. The other special token is a separator token [SEP]. It simply indicates the end of a segment and the beginning of the succeeding segment. The type of segments used depends on the specific task BERT is fine-tuned for. For instance, in question answering one segment can be a question, whereas the other one can be the reference text. The two segments are then appended and separated by a special separator token [SEP]. In our model, we chose the context of the aspect (i.e. the whole sentence) as one segment and the aspect of SA as the other.

The embedding layer shown in Figure 6.1 illustrates the input format that BERT expects. In this example, the sentence and the aspect in question are "This medicine works great when it comes to pain." and "pain", respectively, which are combined into the following input sequence: "[CLS] This medicine works great when it comes to pain . [SEP] pain [SEP]". Finally, to meet the fixed-length requirement that BERT expects of its input, such sequence is padded using a special padding token [PAD] until the maximum length of 70 tokens has been reached.

The input sequence is then processed as follows. First, each token's vocabulary identifier is mapped to a token embedding that was learned during training. A binary vector is then used to differentiate between two text segments. The binary vector is mapped to a segment embedding using a lookup table, which was also learned during training. Next, local token positions are mapped to positional embeddings using a lookup table, which was updated during training. Finally, the attention mask represents an array of 0s and 1s, indicating which tokens are padding and which are not, respectively.

The three types of embeddings, which correspond to tokens, segments and positions respectively, are added up and passed to the pre-trained BERT$_{\text{base}}$ model, along with the attention mask, which comprises 12 layers of transformer encoders, each having a hidden size of 768 and 12 attention heads. Each layer produces a token-specific output, which can then be used as its contextualised embedding. The context-sensitive nature of BERT embeddings makes this language model well suited for the task of aspect-based SA as the embedding of the aspect will account for the words surrounding it. Of note, BERT uses WordPiece tokenization to obtain subword units by applying a greedy segmentation algorithm to minimise the number of WordPieces in the training corpus [262]. Therefore, the embedding of out-of-vocabulary words can be assembled from its subwords.

Similarly to binary classification tasks described originally in [92], the final transformer output that corresponds to the special [CLS] token amounts to an aggregate problem representation, i.e. a pooled output. To determine the sentiment from this aggregate representation of a sentence and its aspect, the pooled output is fed into the classification layer. The classification layer reduces the size of the pooled output to 2 dimensions, which correspond to the log-odds (or logits) of the classification output with respect to the question of whether the implied sentiment is positive or negative. The classification layer was not pre-trained unlike the preceding layers of the neural network. Multiple pre-trained BERT models can be used here. They differ with respect to the choice of hyperparameter values. We employed the BERT$_{\text{base}}$ model, which was pre-trained using 12 layers of transformer encoders, 12 attention heads and the hidden dimension of 768. Going back to the classification layer, its output is passed through the softmax function, which estimates the probability distribution over positive and negative sentiments.

6.2.2 Implementation and training

To implement our approach described above, we used the publicly available pre-trained BERT$_{\text{base}}$ model. Specifically, we used its distribution from Hugging Face, an open-source library which consists of state-of-the art transformer architectures under a unified API [258]. The pre-trained BERT model was fine-tuned by minimising cross-entropy loss (defined in Chapter 4), which is calculated between the output from the softmax and the true labels. The loss function was optimised with Adam optimizer [147], a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments, with the learning rate set to 2x10^{-5}. The specific learning rate was selected based on the suggestions made in [92]. All other hyperparameters were set to their default values.

The classification model was trained for 4 epochs following the recommendations of BERT's original authors to use 2-4 epochs to fine-tune BERT for a specific NLP downstream task [92]. We evaluated the model after each epoch on the validation set. During each epoch, the model parameters were updated with respect to the error of each batch of the training data. Batch size for training, validation and test sets was set to 16.

The overall SA system was implemented in Python programming language using PyTorch [203], a deep learning framework which combines usability and speed by coding executable models, thus making debugging easier, while being efficient and supporting further hardware acceleration. All our experiments were run on a CPU, not a GPU, of a PC with an Intel processor with 6 cores each running at 2.6GHz and 16GB RAM.

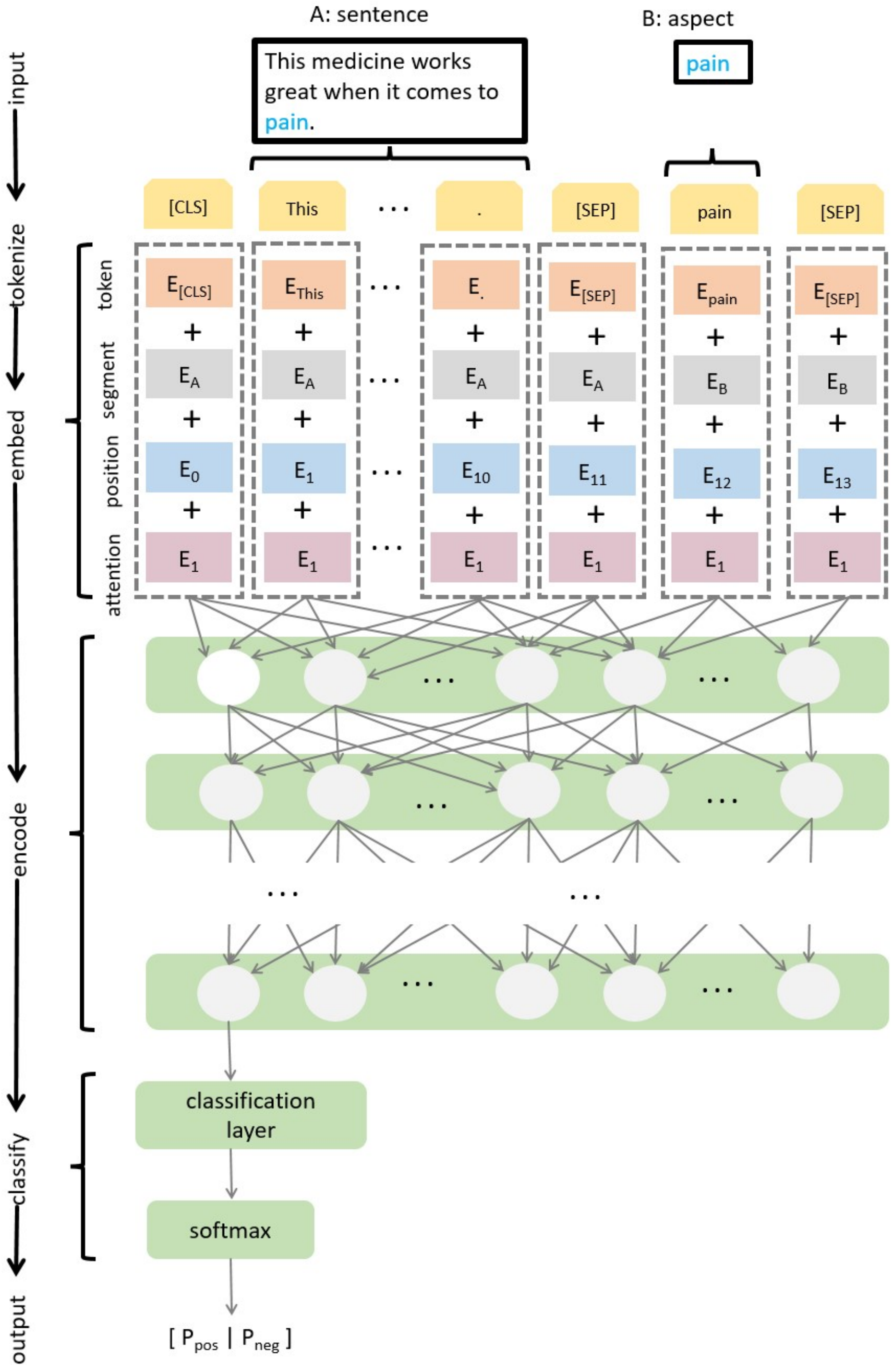

Figure 6.1: BERT-based architecture for aspect-based SA.

6.3 Results

In this Section we evaluate the proposed model for aspect-based SA. We re-used the dataset we described in detail in Section 5.3.1. This allowed us to re-use the graph convolution model described in Chapter 5 as a baseline. The source code is available at https://github.com/zanastazia/ABSA-attention_or_co-dependency.

6.3.1 Evaluation

To be able to make the direct comparison to the baseline results, we re-used the evaluation measures described in the previous Chapter. Specifically, we used measures commonly used to evaluate classification performance including accuracy and cross-entropy loss. Accuracy represents the percentage of correctly classified instances. Accuracy is not always a reliable metric. For example, it may provide misleading results when the test dataset is unbalanced. As we can see from Table 5.3, this is not the case in this study, thus justifying the use of this metric. Given that our model also provides probability distribution over the sentiment labels as output, we used cross-entropy loss to compare the predicted probabilities to the gold standard labels as follows:

$$loss = -\frac{1}{n}\sum_{i=1}^{n} ln(p_i) \tag{6.1}$$

where p_i is the corrected probability, i.e. the probability that a particular prediction matches the gold standard label. The closer the predicted probability to the gold standard label, the lower the cross-entropy loss.

In addition to running experiments using the standard BERT model, we performed experiments with its distilled version. DistilBERT is a smaller general-purpose language model, which can be fine-tuned for specific tasks just like its larger counterpart [223]. It reduces the size of a BERT model by 40% while retaining 97% of its language understanding capabilities with a benefit of being 60% faster to run. Both language models

Method		**Accuracy**	**Loss**
Baseline		81.78%	0.4570
$\text{BERT}_{\text{base}}$	uncased	78.14%	0.5270
	cased	94.33%	0.3641
$\text{DistilBERT}_{\text{base}}$	uncased	73.28%	0.5688
	cased	94.74%	0.3660

Table 6.1: The evaluation results.

come in both cased and uncased versions. In the uncased models, the text is lowercased prior to WordPiece tokenization, thus making the model case insensitive. No case changes are performed on text in the cased version.

Table 6.1 provides the results. The significance of case can be immediately observed. In both BERT and DistilBERT, the cased model outperformed the uncased one by a large margin with the accuracy in the 70s and 90s, respectively. In fact, the uncased models performed worse than the baseline. On the other hand, the cased models not only outperformed the baseline by more than 12 percent points but fell short of the perfect accuracy of 100% by only 5 percent points. The difference in performance between the two cased models was negligible.

6.4 Discussion

The impact of casing on the performance of SA was unexpected, so it warrants further analysis to try to explain this phenomenon. Intuitively, one might expect this issue to be related to the use of the personal pronoun I, which is often used to describe one's state. In particular, within the realm of health and well-being the usage of pronouns was found to have an effect on SA even more so than on standard English usage [193]. The total number of sentences that were correctly classified by the cased model but incorrectly classified by the uncased model was 47. Therefore, the error analysis was

ID	Sentence	Label	Uncased	Cased
1	Excellent **headache** reliever!	$+$	$-$	$+$
2	Good medicine, it gets rid of your **pain** without that drowsy sick feeling.	$+$	$-$	$+$
3	Love this medicine, no **headache**.	$+$	$-$	$+$
4	Sadly no effect on my **pain**.	$-$	$+$	$-$
5	Made my **symptom** worse - so much for 24 hour relief.	$-$	$+$	$-$
6	No **pain** relief whatsoever.	$-$	$+$	$-$

Table 6.2: A sample of sentences incorrectly classified by the uncased model that are correctly classified by the cased model.

not a laborious undertaking. Table 6.2 provides a sample of errors made by the uncased model, which were corrected by the cased model. For simplicity, the results in this Section are based on the standard BERT model.

Within 47 sentences we found only 10 mentions of the personal pronoun I that was not at the beginning of the sentence. In the majority of cases such as those shown in Table 6.2 we can see that the personal pronoun I did not play any role in these sentences, so we can dismiss our initial hypothesis and investigate other possible effects of casing on the classification performance. In the same table, the words highlighted using a bold typeset represent the aspect. The case of an aspect was clearly not affected by lowercasing. In fact, none of the other words were affected by lowercasing apart from the first word of a sentence. English grammar requires the first word of a sentence to be capitalised. A quick inspection of the first words reveals the majority to be emotionally charged words that are typically found in most sentiment lexica, e.g. excellent, good, love and sadly. We inspected all errors and indeed found that in all such cases the correct sentiment of the whole sentence coincided with the sentiment of these words. When the model was pre-trained it was reasonable to assume that these words were also found at the beginning of a sentence as there are few other cases that would require their capitalisation. Therefore, their learnt embeddings would be correlated with their initial position in a sentence. When the model is fine-tuned these words also have the most immediate impact on the neighbouring special token [CLS], which represents a pooled output. Therefore, it is reasonable to assume that the performance of the model is more directly linked to the position of these words rather than their casing alone.

The baseline model was not case sensitive. It also used convolution relative to the aspect of a sentence. It was, therefore, less influenced by the initial word unlike the BERT model that uses a pooled output that is associated with a special token positioned before the start of a sentence. Nonetheless, BERT outperformed the baseline approach. To investigate the internal logic of the BERT model we used Captum [150], an open-source library for model interpretability. It uses integrated gradients [239], an axiomatic attribution method that attributes the prediction of a deep neural network to its inputs. Two fundamental axioms that an attribution method should satisfy ensure that any artefacts affecting the attribution method are related to either the data or the neural network rather than the method itself. The first axiom, sensitivity, states that (1) whenever input and baseline differ in only one feature but have different predictions, then that particular feature should be given a non-zero attribution, and (2) if the function implemented by the neural network does not depend on some variable, then that particular variable should always be given zero attribution. The second axiom, implementation invariance, states that any two functionally equivalent networks should receive identical attributions regardless of any differences in their implementations.

Of note, this attribution method only measures the relative importance of features in a neural network but does not address the interactions between the features nor the internal logic of the network. To study the extent to which syntactic dependencies between an aspect and other tokens (i.e. features in this context) are correlated with the attributions assigned to these tokens we cross-referenced the attribution scores received by each token to its distance from the aspect in the syntactic dependency graph. Figures 6.2-6.6 provide examples of cross-referencing a token's distance from the aspect to its attribution score. The zero distance in the dependency graph, which is provided at the top, indicates an aspect of SA. The attribution scores given at the bottom have been colour-coded using the heatmap colour palette given on the right with lighter colours indicating higher attribution score. Let us have a closer look at these examples. In Figure 6.2, we can observe that the tokens that are one to two steps away from the aspect in the dependency graph received the highest attribution score with the exception of the punctuation token. Of note, the closest token to the aspect's right in the sequence graph (i.e. the word "and") received the lowest score. A similar trend continues to the right including the positive word "help" also receiving a low attribution score and thus not contributing significantly to the positive sentiment of the aspect "pain". Similar trends could be observed in the example shown in Figure 6.3. Even though the positive word "great", being two steps away from the aspect, received a negative attribution score, the sentiment was still correctly classified. The highest attribution score given to the word "treating" as the most directly related to the aspect "pain" in the dependency graph possibly played a pivotal role in the SA. In Figure 6.4 the word "awful" received the highest attribution score, which is consistent with both the distance in the dependency graph and the sentiment polarity. In Figure 6.5, the key word for the negative sentiment "not" received a mediocre attribution score as did other words in the same sentence apart from the preposition "for" together with the punctuation token. In the final example shown in Figure 6.6, the aspect "fatigue" appears in coordination with the word "pain", which received one of the highest attribution scores. This coordination increased the distance of other context words from the aspect in the dependency graph making the attribution scores more difficult to interpret.

To see whether this anecdotal evidence can be generalised, we performed statistical analysis to check whether higher attribution scores are correlated with smaller distances in the dependency graph. We used Pearson correlation coefficient, which is

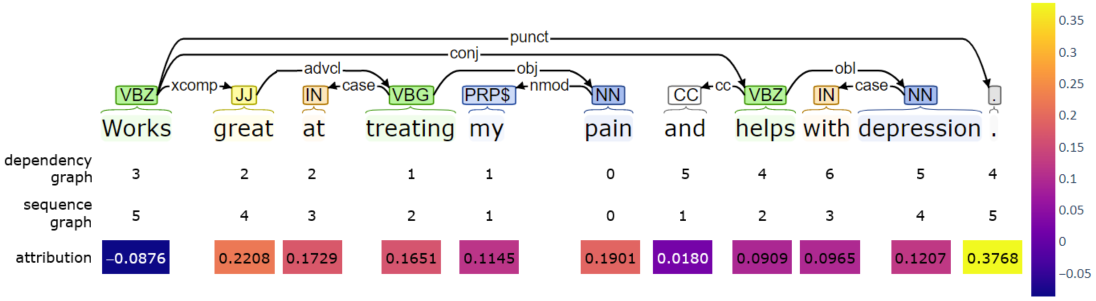
punct
conj
advcl
obj
obl
xcomp
case
nmod
cc
case
VBZ JJ IN VBG PRP$ NN CC VBZ IN NN .
Works great at treating my pain and helps with depression .
dependency graph 3 2 2 1 1 0 5 4 6 5 4
sequence graph 5 4 3 2 1 0 1 2 3 4 5
attribution −0.0876 0.2208 0.1729 0.1651 0.1145 0.1901 0.0180 0.0909 0.0965 0.1207 0.3768
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−0.05

Figure 6.2: An example of "pain" as the aspect of SA with positive polarity.

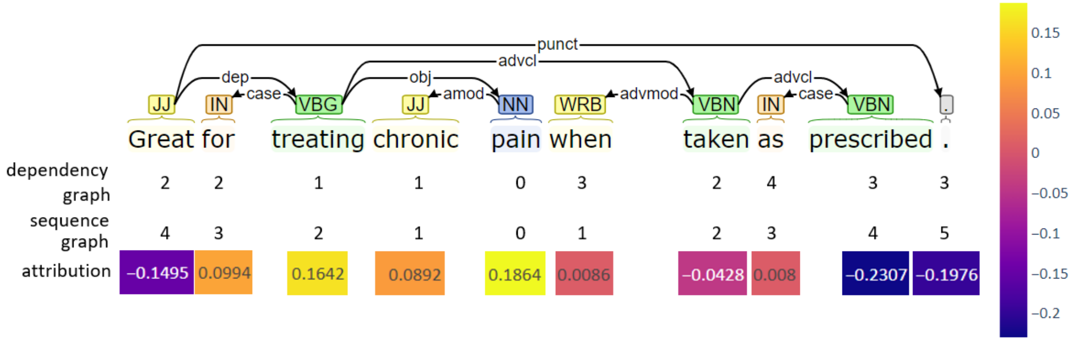
punct
advcl
dep
obj
advcl
case
amod
advmod
case
JJ IN VBG JJ NN WRB VBN IN VBN .
Great for treating chronic pain when taken as prescribed .
dependency graph 2 2 1 1 0 3 2 4 3 3
sequence graph 4 3 2 1 0 1 2 3 4 5
attribution −0.1495 0.0994 0.1642 0.0892 0.1864 0.0086 −0.0428 0.008 −0.2307 −0.1976
0.15
0.1
0.05
0
−0.05
−0.1
−0.15
−0.2

Figure 6.3: An example of "pain" as the aspect of SA with positive polarity.

calculated according to the following formula:

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}, \tag{6.2}$$

where x_i represents the i-th data point in vector x, whereas $\overline{x}$ represents the mean value of vector x. Here, the null hypothesis is that there is no correlation between the

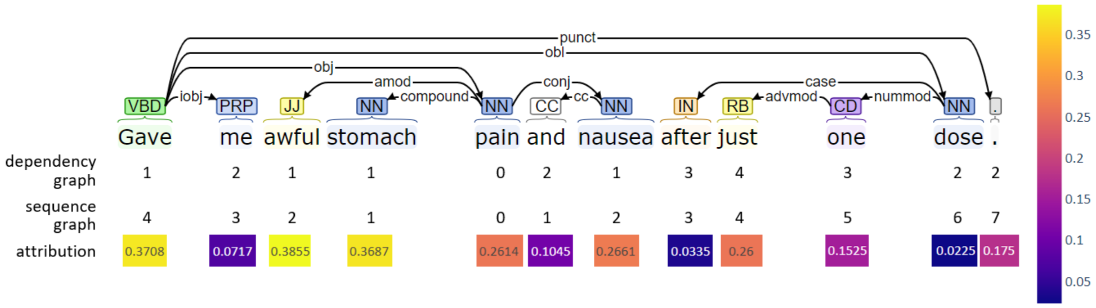
punct
obl
obj
amod
conj
case
iobj
compound
cc
advmod
nummod
VBD PRP JJ NN NN CC NN IN RB CD NN .
Gave me awful stomach pain and nausea after just one dose .
dependency graph 1 2 1 1 0 2 1 3 4 3 2 2
sequence graph 4 3 2 1 0 1 2 3 4 5 6 7
attribution 0.3708 0.0717 0.3855 0.3687 0.2614 0.1045 0.2661 0.0335 0.26 0.1525 0.0225 0.175
0.35
0.3
0.25
0.2
0.15
0.1
0.05

Figure 6.4: An example of "pain" as the aspect of SA with negative polarity.

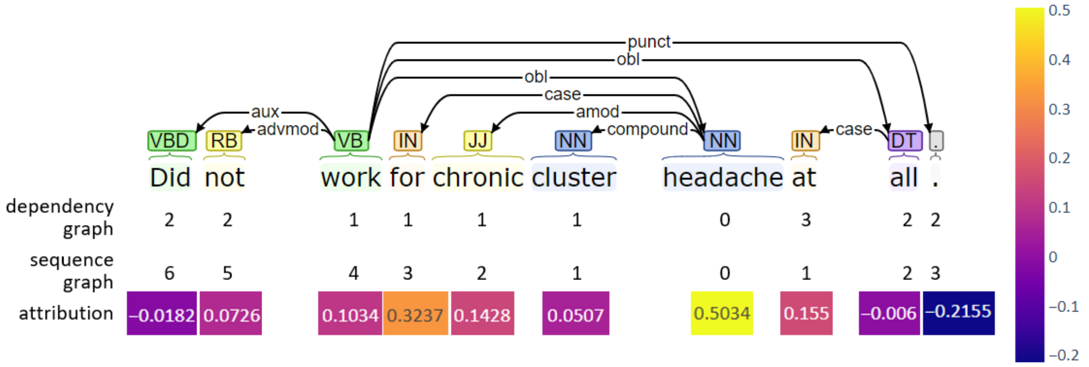
punct
obl
obl
case
amod
compound
aux
advmod
case
VBD RB VB IN JJ NN NN IN DT .
Did not work for chronic cluster headache at all .
dependency graph 2 2 1 1 1 1 0 3 2 2
sequence graph 6 5 4 3 2 1 0 1 2 3
attribution −0.0182 0.0726 0.1034 0.3237 0.1428 0.0507 0.5034 0.155 −0.006 −0.2155
0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2

Figure 6.5: An example of "headache" as the aspect of SA with negative polarity.

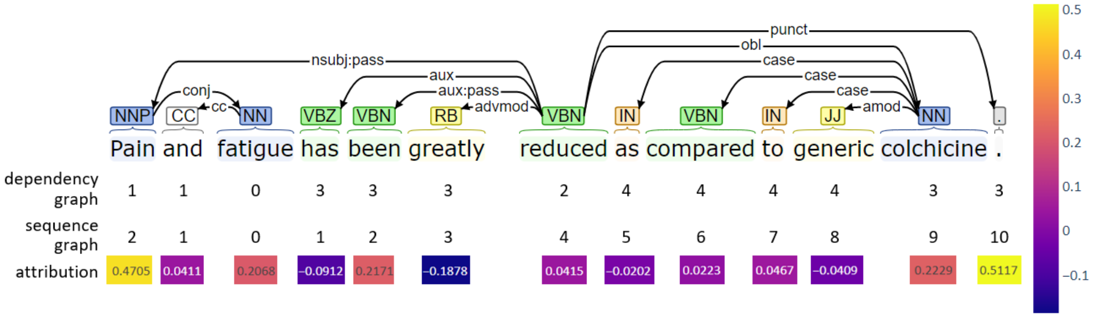
punct
obl
nsubj:pass
case
aux
case
aux:pass
case
conj
cc
advmod
amod
NNP CC NN VBZ VBN RB VBN IN VBN IN JJ NN .
Pain and fatigue has been greatly reduced as compared to generic colchicine .
dependency graph 1 1 0 3 3 3 2 4 4 4 4 3 3
sequence graph 2 1 0 1 2 3 4 5 6 7 8 9 10
attribution 0.4705 0.0411 0.2068 −0.0912 0.2171 −0.1878 0.0415 −0.0202 0.0223 0.0467 −0.0409 0.2229 0.5117
0.5
0.4
0.3
0.2
0.1
0
−0.1

Figure 6.6: An example of "fatigue" as the aspect of SA with positive polarity.

attribution score and the distance of the token from the aspect in the dependency graph. The correlation between the two variables was found to be -0.074. In other words, the smaller the distance the higher attribution score and vice versa. The corresponding p-value of $5.4732 * 10^{-19}$ was smaller than the set threshold of 0.05 indicating that the correlation between the two variables was statistically significant. Therefore, the null hypothesis was rejected.

We performed analogous experiments using the sequence graph representation, i.e. we measured the correlation between the token attribution score and the distance of the token from the aspect in the sequence graph. The correlation between these two variables was found to be -0.069 with the p-value of $1.4107 * 10^{-16}$. It came as no surprise that the local context of an aspect was found to play an important role in determining its sentiment. This could be partly due to an overlap of tokens that are

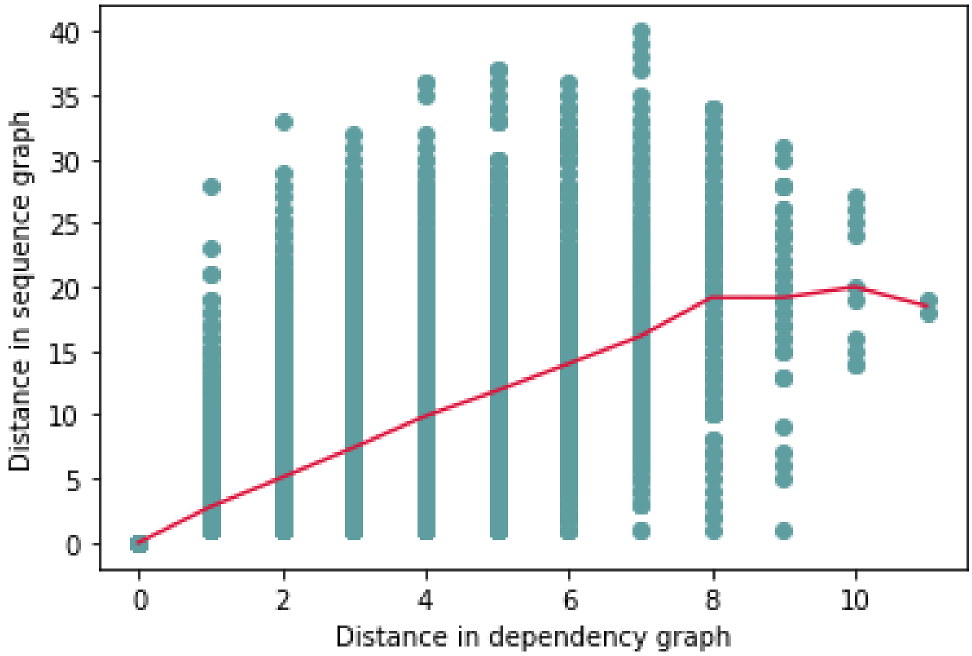

Figure 6.7: Relationship between the distance in dependency graph and distance in sequence graph.

close to the aspect in both dependency and sequence graphs. However, the scatter plot shown in Figure 6.7, which illustrates the relationship between the two ways of measuring the distance from the aspect, indicates that this is not generally the case. For example, tokens that are 2 steps away from the aspect in the dependency graph are on average 5 steps away in the sequence graph.

We further compared the average attribution score against the distance of a token in both representations in Figure 6.8. We can see that the average scores for tokens that are one or two steps away from the aspect do not vary much between the two representations. Interestingly, the attribution score across the sequence graph distances is near constant for all tokens that are between one and six tokens away. On the other side, we can observe a sharp decline in the attribution score for distances over 3 steps away in the dependency graph. This would indicate that the dependency graph distance is a better discriminator of relevant features according to their attribution score. This agrees with the previous finding that the correlation between the token attribution score and the distance of the token from the aspect was stronger for the dependency graph (-0.074) than it was for the sequence graph (-0.069). We, therefore, conclude that the BERT model accounts for syntactic dependencies when performing sentiment classification.

This is in agreement with previous observations that some attention heads approximate syntactic structure by specialising to track individual dependency types [129]. Moreover, individual dependency types are often tracked by the same heads across typologically diverse languages [215]. At the same time, not all dependency types are tracked with the same robustness [76]. Prioritising certain types of dependencies over others may provide a plausible explanation as to why the fine-tuned BERT model outperformed our previous GCN-based approach described in Chapter 5.

Namely, two successive convolutions were performed on each word represented by its embedding following the edges in the syntactic dependency graph, hence propagating information across the graph to the second order neighbour. This approach significantly outperformed alternative approaches, which did not take the syntactic structure into account, hence its success was attributed to the way in which it incorporated syntactic dependencies into the logic of the neural network. However, despite their apparent value for the task of aspect-based SA, the convolution was applied to all syntactic dependencies indiscriminately. In other words, pre-determined convolution across explicit syntactic dependencies. In this study, the test data suggest that the model takes into account implicit syntactic dependencies with added flexibility of varying attention across these dependencies. The flexibility of the transformer-based approach embodied in the attention, which is used to prioritise certain types of information including different dependency types, may hold the key to the superior performance of the transformed-based approach compared to that of the GCN-based one.

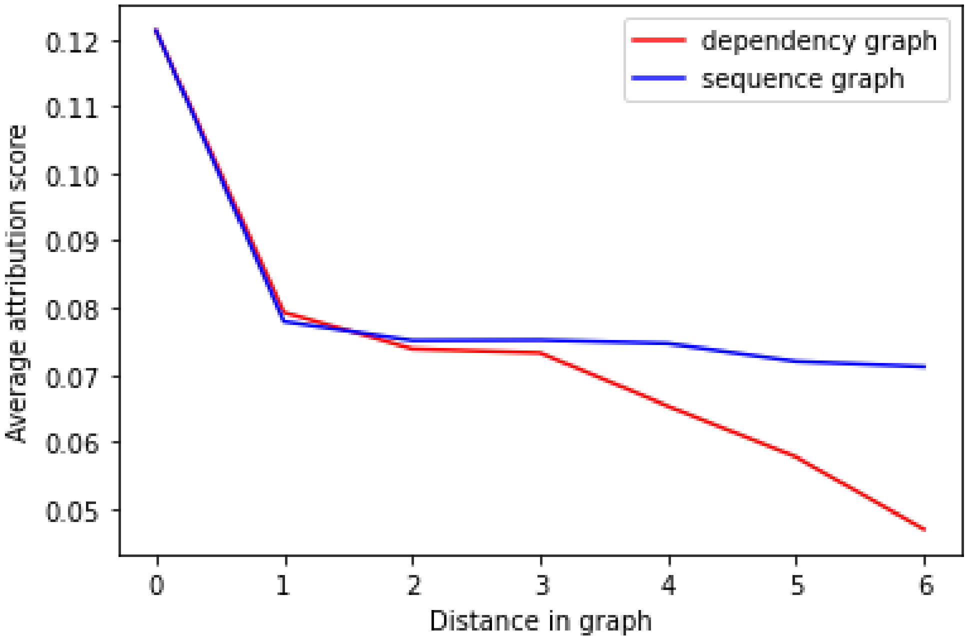

Figure 6.8: Average attribution score against the distance within the dependency and sequence graph.

6.5 Summary

In this Chapter we presented an approach to fine-tuning the BERT language model for the specific task of aspect-based SA. BERT is pre-trained on a large dataset, which makes it robust with respect to the out-of-vocabulary problem and allows for fine-tuning the model for a specific NLP task by using a relatively small dataset. Our fine-tuned model achieved the accuracy of approximately 95% on a well-balanced test set. It outperformed our previous approach which used syntactic information to guide the operation of a neural network. Our latest approach demonstrated that a BERT-based model can not only compensate for the lack of explicit syntactic information but can in fact offer superior performance. Previous studies provided evidence that during the training phase BERT does learn some forms of linguistic structure [129, 215, 76]. In this study, we provide further evidence of this phenomenon in the context of aspect-based SA. Specifically, we focused on the syntactic dependencies that involve a given aspect. The evidence suggests that the model's attention is correlated with the degree of separation from an aspect calculated as the number of steps away from the aspect in a syntactic dependency graph. This correlation was found to be stronger than the one calculated for the distance in the flat sentence representation. This brings us to the conclusion that the BERT model accounts for the syntactic dependencies when classifying the sentiment of the given aspect.

Finally, the high accuracy of the model achieved in the realm of health and well-being opens up an array of possible applications in this domain [233]. When it comes to health, modern society tends to be preoccupied with the inherently negative phenomena such as diseases, injuries and disabilities [56]. However, for chronic patients, achieving a good quality of life does not necessarily imply the absence of symptoms that are associated with their medical condition. In reality, their quality of life is determined by the extent to which these symptoms are effectively managed. However, the negative sentiment associated with health symptoms a priori tends to skew the results of SA toward the negative spectrum. Previously, such a priori bias made it difficult to measure sentiment in this domain as reported in Chapter 2. This Chapter provides evidence that a BERT model can be successfully fine-tuned to overcome this obstacle. The ability to accurately measure the sentiment associated specifically with signs and symptoms can support the development of systems designed to engage patients and monitor their self-management of chronic conditions remotely [232].

Chapter 7

Conclusion

We hereby conclude this thesis by summarising the key findings and outlining the possible ways in which the presented research can be extended.

We embarked on this research upon identifying a gap related to the underperformance of off-the-shelf sentiment analysis (SA) tools in the domain of health and well-being. The research gap was established by performing a systematic literature review of SA in this domain, which was presented in Chapter 2. It highlighted the fact that deep learning has not been routinely applied to support this task to analyse narratives related to this area of life. Given a possibility that such underperformance could have been related to using outdated methods, we were keen to explore the potential of deep learning in bridging this gap.

To investigate other possible causes of such underperformance, we performed a comparative analysis of off-the-shelf tools on a dataset related to health and well-being and described the findings in Chapter 3. We hypothesised that a key factor contributing to the underperformance of SA could be the negative bias associated with the concepts related to health and well-being, which correspond to one's symptoms or conditions, e.g. headache, nausea, pain. To test this hypothesis, we effectively masked such concepts. They were identified in free text by relying on the structure of the Unified Medical Language System, a large repository of inter-related biomedical concepts [58]. By re-running the given set of tools on the masked text, we noticed an improvement in the performance. This reiterated the need for novel methods that can modify the sentiment of such concepts in response to their context.

This has led to the research hypothesis that we could use syntactic dependencies to improve the performance of aspect-based SA in health and well-being. The results of our research confirmed this to be the case. The research itself was structured against the following objectives:

RO1. Finding a suitable problem representation of aspect-based sentiment analysis that takes into account syntactic dependencies.

RO2. Developing a deep learning approach that effectively operates on such problem representation.

RO3. Providing evidence that the actions taken in response to the previous two objectives lead to performance improvements of aspect-based sentiment analysis in health and well-being.

To address RO1, we started by noticing that the sentiment expressed towards specific aspects in written narratives is highly context dependent. For example, not all mentions of someone's symptoms are negative. If we look at the following narrative "I do not feel any pain.", we can see that the negation of "pain" as the aspect of SA indicates a positive outcome and hence expresses a positive sentiment. Negation can be explicitly encoded by dependency parse trees (see Figure 5.4 for an example). Moreover, syntactic dependencies can be used to represent other ways of modifying an aspect.

To address RO2, that is - to utilise syntactic dependencies as features and to enable the model to make the best use of them, in Chapter 5 we proposed a new approach to aspect-based SA based on the graph convolution over the dependency parse trees (for overall system design see Figure 5.1). Convolution was applied by traversing the syntactic dependencies associated with the aspect of SA. More precisely, two successive convolutions were performed on each word represented by its embedding following the edges in the syntactic dependency graph, hence propagating information across the graph to the second order neighbour. The experimental results showed that this approach outperformed other neural network architectures such as RNN and LSTM. It achieved the accuracy of approximately 82% outperforming a RNN which achieved the accuracy of less than 68%. The experiments also suggested that the dependency parse trees make better utilisation of context to perform aspect-based SA when compared to the simple flat representation of the sentence. In this study we successfully tackled the negative bias by exploiting the syntactic dependencies and we effectively closed the performance gap by achieving state-of-the-art results within the domain of health and well-being. Our approach significantly outperformed alternative approaches, which did not take the syntactic structure into account, hence its success was attributed to the way in which it incorporated syntactic dependencies into the logic of the neural network.

To address RO3, that is - to evaluate the newly proposed model, we created a new dataset as there were no publicly available datasets suitable for the evaluation of aspect-based SA in the domain of health and well-being. The dataset consists of drug reviews that were taken from the publicly available dataset [109]. All reviews were automatically annotated with the concepts from the Unified Medical Language System [58], which were treated as the aspect of SA.

In the meantime, the appearance of transformer-based pre-trained language models such as BERT, which provide contextualised word embeddings, has taken the field of NLP to a new level. These developments inspired additional research into the given topic of aspect-based SA the results of which we presented in Chapter 6. Specifically, we fine-tuned the BERT language model for the task of aspect-based SA. This model outperformed our previous convolutional approach, described in Chapter 5, by achieving the accuracy of approximately 95%. To tie these results to our research hypothesis, which states that syntactic dependencies can improve the performance of aspect-based SA, we conducted additional analysis to investigate whether this model integrated any features that can be interpreted as syntactic dependencies.

We found that the distance within the dependency graph was an important discriminator of relevant features. This was in agreement with previous observations that some attention heads approximate syntactic structure by specialising to track individual dependency types [129]. However, not all dependency types are tracked with the same robustness [76]. Prioritising certain types of dependencies over others may provide a plausible explanation as to why the fine-tuned BERT model outperformed our previous GCN-based approach. Despite their apparent value for the task of aspect-based SA, the convolution was applied to all syntactic dependencies indiscriminately. Our analysis indicated that the BERT-based model takes into account implicit syntactic dependencies with added flexibility of varying attention across these dependencies. The flexibility of the transformer-based approach embodied in the attention, which is used to prioritise certain types of information including different dependency types, may hold the key to the superior performance of the transformed-based approach compared to that of the GCN-based one.

In summary, our results from both original studies support the hypothesis that the consideration of syntactic dependencies does improve the performance of aspect-based SA as demonstrated in a particularly challenging domain of health and well-being.

7.1 Limitations and Future Work

The aspect-based SA approaches proposed in Chapter 5 and Chapter 6 assume that the aspects in question are given a priori. This limitation could be addressed in future research by focusing on approaches that identify the aspects of SA automatically. This would also help adapting the proposed approach for other domains, where suitable ontologies are not readily available. It would also enable broader exploration of sublanguages in the context of aspect-based SA with a focus on the a priori sentiment of aspects and the nature of their interaction with the surrounding context. Although it is known that SA suffers from domain dependency [63], the issues causing it certainly warrant further investigation and an attempt to explain this phenomenon using the sublanguage theory of Zellig Harris [117]. This would require development of domain-specific datasets. Some domains, especially those with commercial applications such as customer reviews, already have open-source datasets available for research. While we bootstrapped such a dataset in relation to health and well-being, this domain would benefit from a community-curated large-scale dataset. A conference workshop bringing together the academic, clinical, industrial and patient communities would provide an opportunity to create such resource and facilitate further research in aspect-based SA in health and well-being. Such research may focus on aggregating the sentiment related to a specific aspect across the whole document not just individual sentences as we proposed in this project. Finally, we used pre-trained word embeddings. Further performance improvements could be gained by optimising the embeddings to reflect the underlying sentiment by providing a clear separation between positive and negative words in the vector space.

Bibliography

[1] Affective Norms for English Words (ANEW). https://csea.phhp.ufl.edu/media/anewmessage.html.

[2] AFINN. http://www2.imm.dtu.dk/pubdb/pubs/6010-full.html.

[3] Ask a Patient. https://www.askapatient.com/.

[4] Breast Cancer Information and Support. https://www.breastcancer.org/.

[5] Cancer Survivors Network. https://csn.cancer.org/.

[6] Cochrane Library: Cochrane Review. https://www.cochranelibrary.com/.

[7] DailyStrength. https://www.dailystrength.org/.

[8] DiabetesDaily. https://www.diabetesdaily.com/.

[9] DrugLib.com. http://www.druglib.com/.

[10] Drugs.com. https://www.drugs.com/.

[11] EBSCO Health. CINAHL Database. https://health.ebsco.com/products/the-cinahl-database/.

[12] Embase. https://www.embase.com/.

[13] Harvard General Inquirer. http://www.wjh.harvard.edu/~inquirer/homecat.htm.

[14] Keras. https://keras.io/.

[15] Language Assessment by Mechanical Turk (labMT) Sentiment Words. https://trinker.github.io/qdapDictionaries/labMT.html.

[16] LIBSVM – A Library for Support Vector Machines. https://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[17] MedHelp. https://medhelp.org/.

[18] MPQA. http://mpqa.cs.pitt.edu/.

[19] National Health Service. https://www.nhs.uk/.

[20] National Library of Medicine. MEDLINE: Description of the Database. https://www.nlm.nih.gov/medline/medline_overview.html/.

[21] NRC Emotion Lexicon. https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.

[22] NRC Emotion Lexicon. http://sentiment.nrc.ca/lexicons-for-research/.

[23] Opinion Lexicon. https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon.

[24] OpinionKB. https://doi.org/10.1371/journal.pone.0124993.s001.

[25] PySVM light. http://daoudclarke.github.io/pysvmlight/.

[26] RateMDs. https://www.ratemds.com/.

[27] scikit-learn. https://scikit-learn.org/stable/.

[28] SentiSense Affective Lexicon. http://nlp.uned.es/~jcalbornoz/resources.html.

[29] SentiStrength. http://sentistrength.wlv.ac.uk/.

[30] SentiWordNet. https://github.com/aesuli/SentiWordNet.

[31] SVM light. https://www.cs.cornell.edu/people/tj/svm_light/.

[32] TextBlob. https://textblob.readthedocs.io/en/dev/.

[33] Universal Dependencies. https://universaldependencies.org/.

[34] WebMD. https://www.webmd.com/.

[35] Weka. https://www.cs.waikato.ac.nz/ml/weka/.

[36] WordNet Affect. http://wndomains.fbk.eu/wnaffect.html.

[37] WordNet Domains. WordNet-Affect. https://wndomains.fbk.eu/wnaffect.html.

[38] World Health Organisation. Geneva, Switzerland: Constitution of the World Health Organisation, 2006. https://www.who.int/governance/eb/who_constitution_en.pdf/.

[39] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, abs/1603.04467, 2016.

[40] David Z. Adams, Richard Gruss, and Alan S. Abrahams. Automated discovery of safety and efficacy concerns for joint & muscle pain relief treatments from online reviews. *International Journal of Medical Informatics*, 100:108–120, 2017.

[41] Adrian Ahne, Francisco Orchard, Xavier Tannier, Camille Perchoux, Beverley Balkau, Sherry Pagoto, Jessica Lee Harding, Thomas Czernichow, and Guy Fagherazzi. Insulin pricing and other major diabetes-related concerns in the USA: a study of 46407 tweets between 2017 and 2019. *BMJ Open Diabetes Research and Care*, 8(1):e001190, 2020.

[42] Altug Akay, Andrei Dragomir, and Björn-Erik Erlandsson. A novel data-mining approach leveraging social media to monitor consumer opinion of sitagliptin. *IEEE Journal of Biomedical and Health Informatics*, 19(1):389–396, 2013.

[43] Ahmet Emre Aladağ, Serra Muderrisoglu, Naz Berfu Akbas, Oguzhan Zahmacioglu, and Haluk O. Bingol. Detecting suicidal ideation on forums: proof-of-concept study. *Journal of Medical Internet Research*, 20(6):e9840, 2018.

[44] Farrokh Alemi and Harry Jasper. An alternative to satisfaction surveys: let the patients talk. *Quality Management in Healthcare*, 23(1):10–19, 2014.

[45] Farrokh Alemi, Manabu Torii, Laura Clementz, and David C. Aron. Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments. *Quality Management in Healthcare*, 21(1):9–19, 2012.

[46] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, MN, USA, 2019. Association for Computational Linguistics.

[47] Muhammad Zubair Asghar, Shakeel Ahmad, Maria Qasim, Syeda Rabail Zahra, and Fazal Masud Kundi. SentiHealth: creating health-related sentiment lexicon using hybrid approach. *SpringerPlus*, 5(1):1–23, 2016.

[48] Muhammad Zubair Asghar, Aurangzeb Khan, Shakeel Ahmad, Maria Qasim, and Imran Ali Khan. Lexicon-enhanced sentiment analysis framework using rule-based classification scheme. *PLOS ONE*, 12(2):e0171649, 2017.

[49] Giuseppe Attardi and Maria Simi. Blog mining through opinionated words. In *Proceedings of TREC 2006, the Fifteenth Text Retrieval Conference*. NIST, January 2006.

[50] İrfan Aygün, Buket Kaya, and Mehmet Kaya. Aspect based Twitter sentiment analysis on vaccination and vaccine types in COVID-19 pandemic with deep learning. *IEEE Journal of Biomedical and Health Informatics*, 2021.

[51] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *LREC 2010, The Seventh International Conference on Language Resources and Evaluation*, pages 2200–2204, Valletta, Malta, 2010. European Language Resources Association (ELRA).

[52] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *CoRR*, abs/1803.01271, 2018.

[53] Xuefeng Bai, Pengbo Liu, and Yue Zhang. Exploiting typed syntactic dependencies for targeted sentiment classification using graph attention neural network. *CoRR*, abs/2002.09685, 2020.

[54] Lingxian Bao, Patrik Lambert, and Toni Badia. Attention and lexicon regularized lstm for aspect-based sentiment analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 253–259, Florence, Italy, 2019. Association for Computational Linguistics.

[55] Mohammad Ehsan Basiri, Shahla Nemati, Moloud Abdar, Erik Cambria, and U. Rajendra Acharya. ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. *Future Generation Computer Systems*, 115:279–294, 2021.

[56] Ole Berg. Health and quality of life. *Acta Sociologica*, 18(1):3–22, 1975.

[57] Jiang Bian, Yunpeng Zhao, Ramzi G. Salloum, Yi Guo, Mo Wang, Mattia Prosperi, Hansi Zhang, Xinsong Du, Laura J. Ramirez-Diaz, Zhe He, and Yuan

Sun. Using social media data to understand the impact of promotional information on laypeopleâs discussions: a case study of Lynch syndrome. *Journal of Medical Internet Research*, 19(12):e414, 2017.

[58] Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270, 2004.

[59] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[60] Margaret M. Bradley and Peter J. Lang. Affective Norms for English Words (ANEW): Instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, 1999.

[61] Ngot Bui, John Yen, and Vasant Honavar. Temporal causality analysis of sentiment change in a cancer survivor network. *IEEE Transactions on Computational Social Systems*, 3(2):75–87, 2016.

[62] Mark L. Cabling, Jeanine W. Turner, Alejandra Hurtado-de Mendoza, Yihong Zhang, Xinyang Jiang, Fabrizio Drago, and Vanessa B. Sheppard. Sentiment analysis of an online breast cancer support group: communicating about Tamoxifen. *Health Communication*, 33(9):1158–1165, 2018.

[63] Erik Cambria, Soujanya Poria, Alexander Gelbukh, and Mike Thelwall. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74–80, 2017.

[64] Xiaodong Cao, Piers MacNaughton, Zhengyi Deng, Jie Yin, Xi Zhang, and Joseph G. Allen. Using Twitter to better understand the spatiotemporal patterns of public sentiment: a case study in Massachusetts, USA. *International Journal of Environmental Research and Public Health*, 15(2):250, 2018.

[65] Jorge Carrillo-de-Albornoz, Ahmet Aker, Emina Kurtic, and Laura Plaza. Beyond opinion classification: Extracting facts, opinions and experiences from health forums. *PLOS ONE*, 14(1):e0209961, 2019.

[66] Jorge Carrillo-de Albornoz, Laura Plaza, and Pablo Gervás. SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis. In *LREC 2012, The Eighth International Conference on Language Resources and Evaluation*, pages 3562–3567, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

[67] Jorge Carrillo-de-Albornoz, Javier Rodriguez Vidal, and Laura Plaza. Feature engineering for sentiment analysis in e-health forums. *PLOS ONE*, 13(11):e0207996, 2018.

[68] Danqi Chen and Christopher D. Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 740–750, Doha, Qatar, 2014. Association for Computational Linguistics.

[69] Long Chen, Xinyi Lu, Jianbo Yuan, Joyce Luo, Jiebo Luo, Zidian Xie, Dongmei Li, et al. A social media study on the associations of flavored electronic cigarettes with health symptoms: observational study. *Journal of Medical Internet Research*, 22(6):e17496, 2020.

[70] Lushi Chen, Tao Gong, Michal Kosinski, David Stillwell, and Robert L. Davidson. Building a profile of subjective well-being for social media users. *PLOS ONE*, 12(11):e0187278, 2017.

[71] Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 452–461, Copenhagen, Denmark, 2017.

[72] Zhipeng Chen and Daniel D. Zeng. Mining online e-liquid reviews for opinion polarities about e-liquid features. *BMC Public Health*, 17(1):1–7, 2017.

[73] Colin Cherry, Saif M. Mohammad, and Berry De Bruijn. Binary classifiers and latent sequence models for emotion detection in suicide notes. *Biomedical Informatics Insights*, 5(Suppl. 1):147–154, 2012.

[74] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, Doha, Qatar, 2014. Association for Computational Linguistics.

[75] Kar-Hai Chu and Thomas W. Valente. How different countries addressed the sudden growth of e-cigarettes in an online tobacco control community. *BMJ Open*, 5(5):e007654, 2015.

[76] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? An analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy, 2019. Association for Computational Linguistics.

[77] Nathan K. Cobb, Darren Mays, and Amanda L. Graham. Sentiment analysis to determine the impact of online messages on smokers' choices to use varenicline. *Journal of the National Cancer Institute Monographs*, 2013(47):224–230, 2013.

[78] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[79] Amy M. Cohn, Michael S. Amato, Kang Zhao, Xi Wang, Sarah Cha, Jennifer L. Pearson, George D. Papandonatos, and Amanda L. Graham. Discussions of

alcohol use in an online social network for smoking cessation: analysis of topics, sentiment, and social network centrality. *Alcoholism: Clinical and Experimental Research*, 43(1):108–114, 2019.

[80] Cristóbal Colón-Ruiz and Isabel Segura-Bedmar. Comparing deep learning architectures for sentiment analysis on drug reviews. *Journal of Biomedical Informatics*, 110:103539, 2020.

[81] Courtney D. Corley, Rada Mihalcea, Armin R. Mikler, and Antonio P. Sanfilippo. Predicting individual affect of health interventions to reduce HPV prevalence. In *Software Tools and Algorithms for Biological Systems. Advances in Experimental Medicine and Biology, volume 696*, pages 181–190. Springer, 2011.

[82] Raminta Daniulaityte, Lu Chen, Francois R. Lamy, Robert G. Carlson, Krishnaprasad Thirunarayan, and Amit Sheth. "When 'bad' is 'good'": identifying personal communication and sentiment in drug-related tweets. *JMIR Public Health and Surveillance*, 2(2):e6327, 2016.

[83] Kia Dashtipour, Soujanya Poria, Amir Hussain, Erik Cambria, Ahmad Y.A. Hawalah, Alexander Gelbukh, and Qiang Zhou. Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive Computation*, 8(4):757–771, 2016.

[84] Kushal Dave, Steve Lawrence, and David M Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *WWW'03: Proceedings of the 12th International Conference on World Wide Web*, pages 519–528, Budapest, Hungary, 2003.

[85] Matthew A. Davis, Kai Zheng, Yang Liu, and Helen Levy. Public response to obamacare on Twitter. *Journal of Medical Internet Research*, 19(5):e167, 2017.

[86] Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Joakim Ginter, Filip andf Nivre, and Christopher D. Manning. Universal

Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 4585–4592, Reykjavik, Iceland, 2014. European Language Resources Association.

[87] Marie-Catherine De Marneffe and Christopher D. Manning. Stanford typed dependencies manual. Technical report, Stanford University, 2008.

[88] Dina Demner-Fushman, Willie J. Rogers, and Alan R. Aronson. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. *Journal of the American Medical Informatics Association*, 24(4):841–844, 2017.

[89] Kerstin Denecke and Yihan Deng. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial Intelligence in Medicine*, 64(1):17–27, 2015.

[90] Bart Desmet and Véronique Hoste. Combining lexico-semantic features for emotion classification in suicide notes. *Biomedical Informatics Insights*, 5(Suppl. 1):125–128, 2012.

[91] Lara L. Devgan, Elizabeth J. Klein, Stephen Fox, and Tugce Ozturk. Bifurcation of patient reviews: An analysis of trends in online ratings. *Plastic and Reconstructive Surgery Global Open*, 8(4):e2781, 2020.

[92] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN, USA, 2019. Association for Computational Linguistics.

[93] Caroline Diorio, Michael Afanasiev, Kristen Salena, and Stacey Marjerrison. 'A world of competing sorrows': A mixed methods analysis of media reports of children with cancer abandoning conventional treatment. *PLOS ONE*, 13(12):e0209738, 2018.

[94] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLOS ONE*, 6(12):e26752, 2011.

[95] Kristina Doing-Harris, Danielle L. Mowery, Chrissy Daniels, Wendy W. Chapman, and Mike Conway. Understanding patient satisfaction with received healthcare services: A natural language processing approach. In *AMIA Annual Symposium Proceedings*, volume 2016, pages 524–533, Chicago, IL, USA, 2016. American Medical Informatics Association.

[96] Eduard C. Dragut, Clement Yu, Prasad Sistla, and Weiyi Meng. Construction of a sentimental word dictionary. In *CIKM '10: Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1761–1764, Toronto, ON, Canada, 2010. ACM.

[97] Jingcheng Du, Jun Xu, Hsing-Yi Song, and Cui Tao. Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data. *BMC Medical Informatics and Decision Making*, 17(2):63–70, 2017.

[98] Jingcheng Du, Jun Xu, Hsingyi Song, Xiangyu Liu, and Cui Tao. Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets. *Journal of Biomedical Semantics*, 8(1):1–7, 2017.

[99] Weifu Du, Songbo Tan, Xueqi Cheng, and Xiaochun Yun. Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. In *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*, pages 111–120, New York, NY, USA, 2010.

[100] Fabon Dzogang, Marie-Jeanne Lesot, Maria Rifqi, and Bernadette Bouchon-Meunier. Early fusion of low level features for emotion mining. *Biomedical Informatics Insights*, 5(Suppl. 1):129–136, 2012.

[101] Miles Efron. Cultural orientation: Classifying subjective documents by cociation analysis. In *Style and Meaning in Language, Art, Music, and Design, Papers from the 2004 AAAI Fall Symposium, Technical Report (7)*, pages 41–48, Arlington, VA, USA, 2004. AAAI Press.

[102] Federico Errica, Marco Podda, Davide Bacciu, and Alessio Micheli. A fair comparison of graph neural networks for graph classification. In *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*, Addis Ababa, Ethiopia, 2020. OpenReview.net.

[103] Carol Friedman, Pauline Kra, and Andrey Rzhetsky. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35(4):222–235, 2002.

[104] Elia Gabarron, Enrique Dorronzoro, Octavio Rivera-Romero, and Rolf Wynn. Diabetes on Twitter: A sentiment analysis. *Journal of Diabetes Science and Technology*, 13(3):439–444, 2019.

[105] Alejandro Garcia-Rudolph, Sara Laxe, Joan Saurí, Montserrat Bernabeu Guitart, et al. Stroke survivors on Twitter: sentiment and topic analysis from a gender perspective. *Journal of medical Internet research*, 21(8):e14077, 2019.

[106] Guido Giunti, Maëlick Claes, Enrique Dorronzoro Zubiete, Octavio Rivera-Romero, and Elia Gabarron. Analysing sentiment and topics related to multiple sclerosis on Twitter. In *30th Medical Informatics Europe conference (MIE 2020)*, pages 911–915, Geneva, Switzerland, 2020. IOS Press.

[107] Sunir Gohil, Sabine Vuik, and Ara Darzi. Sentiment analysis of health care tweets: Review of the methods used. *JMIR Public Health and Surveillance*, 4(2):e5789, 2018.

[108] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[109] Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In *Proceedings of the 2018 International Conference on Digital Health*, pages 121–125, Lyon, France, 2018. ACM.

[110] Scott Gray, Alec Radford, and Diederik P. Kingma. GPU kernels for block-sparse weights. 2017.

[111] Felix Greaves, Daniel Ramirez-Cano, Christopher Millett, Ara Darzi, and Liam Donaldson. Use of sentiment analysis for capturing patient experience from free-text comments posted online. *Journal of Medical Internet Research*, 15(11):e239, 2013.

[112] Oliver Gruebner, Sarah R. Lowe, Martin Sykora, Ketan Shankardass, S. V. Subramanian, and Sandro Galea. A novel surveillance approach for disaster mental health. *PLOS ONE*, 12(7):e0181233, 2017.

[113] Gabrielle Gurdin, Jorge A. Vargas, Luke G. Maffey, Amy L. Olex, Nastassja A. Lewinski, and Bridget T. McInnes. Analysis of inter-domain and cross-domain drug review polarity classification. In *AMIA Summits on Translational Science Proceedings*, volume 2020, pages 201–210, Online, Houston, TX, USA, 2020. American Medical Informatics Association.

[114] Pari Delir Haghighi, Yong-Bin Kang, Rachelle Buchbinder, Frada Burstein, and Samuel Whittle. Investigating subjective experience and the influence of weather among individuals with fibromyalgia: a content analysis of Twitter. *JMIR Public Health and Surveillance*, 3(1):e6344, 2017.

[115] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NIPS 2017)*, volume 30, pages 1024–1034, Long Beach, CA, USA, 2017. Curran Associates Inc.

[116] Yue Han, Meiling Liu, and Weipeng Jing. Aspect-level drug reviews sentiment analysis based on double bigru and knowledge transfer. *IEEE Access*, 8:21314–21325, 2020.

[117] Zellig Harris. *Theory of Language and Information: A Mathematical Approach*. Clarendon Press, 1991.

[118] Kamber L Hart, Roy H Perlis, and Thomas H McCoy Jr. What do patients learn about psychotropic medications on the web? a natural language processing study. *Journal of affective disorders*, 260:366–371, 2020.

[119] Ahmed Hassan and Dragomir Radev. Identifying text polarity using random walks. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 395–403, Uppsala, Sweden, 2010.

[120] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181, Madrid, Spain, 1997. Association for Computational Linguistics.

[121] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *WWW'16: Proceedings of the 25th International Conference on World Wide Web*, pages 507–517, Montréal, Canada, 2016. ACM.

[122] Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. Aspect-based sentiment analysis using bert. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 187–196, 2019.

[123] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, and Jürgen Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In *A Field Guide to Dynamical Recurrent Neural Networks*, pages 237–243. IEEE Press, 2001.

[124] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[125] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.

[126] Anthony M. Hopper and Maria Uriyo. Using sentiment analysis to review patient satisfaction data located on the internet. *Journal of Health Organization and Management*, 29(2):221–233, 2015.

[127] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[128] Yulin Hswen, Amanda Zhang, Kara C. Sewalk, Gaurav Tuli, John S. Brownstein, and Jared B. Hawkins. Investigation of geographic and macrolevel variations in LGBTQ patient experiences: longitudinal social media analysis. *Journal of Medical Internet Research*, 22(7):e17087, 2020.

[129] Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R. Bowman. Do attention heads in BERT track syntactic dependencies? *CoRR*, abs/1911.12246, 2019.

[130] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *KDD'04: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, Seattle, WA, USA, 2004.

[131] Minqing Hu and Bing Liu. Mining opinion features in customer reviews. In *AAAI'04: Proceedings of the 19th National Conference on Artificial Intelligence*, volume 4, pages 755–760, San Jose, CA, USA, 2004.

[132] Binxuan Huang, Yanglan Ou, and Kathleen M. Carley. Aspect level sentiment classification with attention-over-attention neural networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 197–206, Washington DC, USA, 2018. Springer.

[133] Robin Huang, Na Liu, Mary Ann Nicdao, Mary Mikaheal, Tanya Baldacchino, Annabelle Albeos, Kathy Petoumenos, Kamal Sud, and Jinman Kim. Emotion sharing in remote patient monitoring of patients with chronic kidney disease. *Journal of the American Medical Informatics Association*, 27(2):185–193, 2020.

[134] Machteld Huber, J André Knottnerus, Lawrence Green, Henriëtte Van Der Horst, Alejandro R. Jadad, Daan Kromhout, Brian Leonard, Kate Lorig, Maria Isabel Loureiro, Jos W.M. Van Der Meer, Paul Schnabel, Richard Smith, Chris van Weel, and Henk Smid. How should we define health? *The BMJ 2011*, 343:d4163, 2011.

[135] Jina Huh, Meliha Yetisgen-Yildiz, and Wanda Pratt. Text classification for assisting moderators in online health communities. *Journal of Biomedical Informatics*, 46(6):998–1005, 2013.

[136] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International AAAI Conference on Web and Social Media*, volume 8, pages 216–225, Ann Arbor, MICH USA, 2014.

[137] Md. Rafiqul Islam, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M. Kamal, Hua Wang, and Anwaar Ulhaq. Depression detection from social network data using machine learning techniques. *Health Information Science and Systems*, 6(1):1–12, 2018.

[138] Rie Johnson and Tong Zhang. Supervised and semi-supervised text categorization using LSTM for region embeddings. In *ICML'16: Proceedings of the 33rd*

International Conference on Machine Learning, volume 48, pages 526–534, New York, NY, USA, 2016. JMLR.org.

[139] Hyesil Jung, Hyeoun-Ae Park, and Tae-Min Song. Ontology-based approach to social data sentiment analysis: detection of adolescent depression signals. *Journal of Medical Internet Research*, 19(7):e259, 2017.

[140] Daniel Jurafsky and James H. Martin. Constituency parsing. In *Speech and Language Processing (3rd Edition)*, chapter 13, pages 232–245. draft, 2019.

[141] Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten De Rijke. Using WordNet to measure semantic orientations of adjectives. In *LREC 2004, The Fourth International Conference on Language Resources and Evaluation*, volume 4, pages 1115–1118, Lisbon, Portugal, 2004. European Language Resources Association (ELRA).

[142] Yin Kang, Youfa Wang, Dongsong Zhang, and Lina Zhou. The public's opinions on a new school meals policy for childhood obesity prevention in the US: A social media analytics approach. *International Journal of Medical Informatics*, 103:83–88, 2017.

[143] Gopal K. Kanji. *100 Statistical Tests*. Sage, 2006.

[144] Jacqueline Kazmaier and Jan H. van Vuuren. The power of ensemble learning in sentiment analysis. *Expert Systems with Applications*, 187:115819, 2022.

[145] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367–1373, 2004.

[146] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751, Doha, Qatar, 2014. Association for Computational Linguistics.

[147] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR), Conference Track Proceedings*, San Diego, CA, USA, 2015.

[148] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France, 2017.

[149] Barbara Kitchenham. Procedures for performing systematic reviews. *Keele University, Keele, UK*, 33(2004):1–26, 2004.

[150] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for PyTorch. *CoRR*, abs/2009.07896, 2020.

[151] Ioannis Korkontzelos, Azadeh Nikfarjam, Matthew Shardlow, Abeed Sarker, Sophia Ananiadou, and Graciela H Gonzalez. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *Journal of Biomedical Informatics*, 62:148–158, 2016.

[152] Akshi Kumar and Arunima Jaiswal. Systematic literature review of sentiment analysis on Twitter using soft computing techniques. *Concurrency and Computation: Practice and Experience*, 32(1):e5107, 2020.

[153] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[154] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[155] Xin Li, Lidong Bing, Wai Lam, and Bei Shi. Transformation networks for target-oriented sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 946–956, Melbourne, Australia, 2018. Association for Computational Linguistics.

[156] Xin Li and Wai Lam. Deep multi-task learning for aspect term extraction with memory interaction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2886–2892, Copenhagen, Denmark, 2017. Association for Computational Linguistics.

[157] Xinlong Li, Xingyu Fu, Guangluan Xu, Yang Yang, Jiuniu Wang, Li Jin, Qing Liu, and Tianyuan Xiang. Enhancing BERT representation with context-aware embedding for aspect-based sentiment analysis. *IEEE Access*, 8:46868–46876, 2020.

[158] Maria Liakata, Jee-Hyub Kim, Shyamasree Saha, Janna Hastings, and Dietrich Rebholz-Schuhmann. Three hybrid classifiers for the detection of emotions in suicide notes. *Biomedical Informatics Insights*, 5(Suppl. 1):175–184, 2012.

[159] Bin Liang, Hang Su, Lin Gui, Erik Cambria, and Ruifeng Xu. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowledge-Based Systems*, 235:107643, 2022.

[160] Sunghoon Lim, Conrad S. Tucker, and Soundar Kumara. An unsupervised machine learning model for discovering latent infectious diseases using social media data. *Journal of Biomedical Informatics*, 66:82–94, 2017.

[161] Bing Liu. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition*, pages 627–666. Taylor and Francis Group, Boca, 2010.

[162] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.

[163] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Recurrent neural network for text classification with multi-task learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 2873–2879, New York, NY, USA, 2016. AAAI Press.

[164] Rui Liu, Xiaoli Zhang, and Hao Zhang. Web-video-mining-supported workflow modeling for laparoscopic surgeries. *Artificial Intelligence in Medicine*, 74:9–20, 2016.

[165] Sisi Liu and Ickjai Lee. Extracting features with medical sentiment lexicon and position encoding for drug reviews. *Health information science and systems*, 7(1):1–10, 2019.

[166] Christos Livas, Konstantina Delli, and Nikolaos Pandis. "My Invisalign experience": content, metrics and comment sentiment analysis of the most popular patient testimonials on YouTube. *Progress in Orthodontics*, 19(1):1–8, 2018.

[167] Edward Loper and Steven Bird. NLTK: the natural language toolkit. *arXiv CoRR*, cs.CL/0205028, 2002.

[168] Xinyi Lu, Long Chen, Jianbo Yuan, Joyce Luo, Jiebo Luo, Zidian Xie, Dongmei Li, et al. User perceptions of different electronic cigarette flavors on social media: observational study. *Journal of medical Internet research*, 22(6):e17280, 2020.

[169] Yingjie Lu, Yang Wu, Jingfang Liu, Jia Li, and Pengzhu Zhang. Understanding health care social media use from different stakeholder perspectives: a content analysis of an online health community. *Journal of Medical Internet Research*, 19(4):e7087, 2017.

[170] Yue Lu, Malu Castellanos, Umeshwar Dayal, and ChengXiang Zhai. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *WWW'11: Proceedings of the 20th International Conference on World Wide Web*, pages 347–356, Hyderabad, India, 2011. ACM.

[171] Siyin Luo, Youjian Gu, Xingxing Yao, and Wei Fan. Research on text sentiment analysis based on neural network and ensemble learning. *Revue d'Intelligence Artificielle*, 35(1):63–70, 2021.

[172] Xiao Luo, Gregory Zimet, and Setu Shah. A natural language processing framework to analyse the opinions on HPV vaccination reflected in Twitter over 10 years (2008-2017). *Human vaccines & immunotherapeutics*, 15(7-8):1496–1504, 2019.

[173] Kim Luyckx, Frederik Vaassen, Claudia Peersman, and Walter Daelemans. Fine-grained emotion detection in suicide notes: A thresholding approach to multi-label classification. *Biomedical Informatics Insights*, 5(Suppl. 1):61–69, 2012.

[174] Jennifer R. Mammen, James J. Java, Hyekyun Rhee, Arlene M. Butz, Jill S. Halterman, and Kimberly Arcoleo. Mixed-methods content and sentiment analysis of adolescentsâ voice diaries describing daily experiences with asthma and self-management decision-making. *Clinical & Experimental Allergy*, 49(3):299–307, 2019.

[175] 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, Baltimore, Maryland, 2014. Association for Computational Linguistics.

[176] James A. McCart, Dezon K. Finch, Jay Jarman, Edward Hickling, Jason D. Lind, Matthew R. Richardson, Donald J. Berndt, and Stephen L. Luther. Using ensemble models to classify the sentiment expressed in suicide notes. *Biomedical Informatics Insights*, 5(Suppl. 1):77–85, 2012.

[177] Allison A. Merz, Alba Gutiérrez-Sacristán, Deborah Bartz, Natalie E. Williams, Ayotomiwa Ojo, Kimberly M. Schaefer, Melody Huang, Chloe Y. Li,

Raquel Sofia Sandoval, Sonya Ye, Ann M. Cathcart, Anabel Starosta, and Paul Avillach. Population attitudes toward contraceptive methods over time on a social media platform. *American Journal of Obstetrics and Gynecology*, 224(6):597e1–597e14, 2021.

[178] Omar Metwally, Seth Blumberg, Uri Ladabaum, and Sidhartha R. Sinha. Using social media to characterize public sentiment toward medical interventions commonly used for cancer screening: an observational study. *Journal of Medical Internet Research*, 19(6):e200, 2017.

[179] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*, Scottsdale, Arizona, USA, 2013.

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

[181] Shervin Minaee, Elham Azimi, and AmirAli Abdolrashidi. Deep-sentiment: Sentiment analysis using ensemble of cnn and bi-lstm models. *arXiv preprint arXiv:1904.04206*, 2019.

[182] Mark V. Mishra, Michele Bennett, Armon Vincent, Olivia T. Lee, Costas D. Lallas, Edouard J. Trabulsi, Leonard G. Gomella, Adam P. Dicker, and Timothy N. Showalter. Identifying barriers to patient acceptance of active surveillance: content analysis of online patient communications. *PLOS ONE*, 8(9):e68563, 2013.

[183] François Modave, Yunpeng Zhao, Janice Krieger, Zhe He, Yi Guo, Jinhai Huo, Mattia Prosperi, and Jiang Bian. Understanding perceptions and attitudes in breast cancer discussions on Twitter. *Studies in Health Technology and Informatics*, 264:1293–1297, 2019.

[184] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[185] Martin M. Müller and Marcel Salathé. Crowdbreaks: tracking health trends using public social media data and crowdsourcing. *Frontiers in Public Health*, 7, article 81, 2019.

[186] Khalid Nawab, Gretchen Ramsey, and Richard Schreiber. Natural language processing to extract meaningful information from patient experience feedback. *Applied Clinical Informatics*, 11(02):242–252, 2020.

[187] Finn Årup Nielsen. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, volume 718 of *CEUR Workshop Proceedings*, pages 93–98, Heraklion, Crete, Greece, 2011. CEUR-WS.

[188] Azadeh Nikfarjam, Ehsan Emadzadeh, and Graciela Gonzalez. A hybrid system for emotion extraction from suicide notes. *Biomedical Informatics Insights*, 5(Suppl. 1):165–174, 2012.

[189] Samira Noferesti and Mehrnoush Shamsfard. Resource construction and evaluation for indirect opinion mining of drug reviews. *PLOS ONE*, 10(5):e0124993, 2015.

[190] Samira Noferesti and Mehrnoush Shamsfard. Using linked data for polarity classification of patientsâ experiences. *Journal of Biomedical Informatics*, 57:6–19, 2015.

[191] Daniel Noll, Brendan Mahon, Bhavna Shroff, Caroline Carrico, and Steven J. Lindauer. Twitter analysis of the orthodontic patient experience with braces vs Invisalign. *The Angle Orthodontist*, 87(3):377–383, 2017.

[192] Punith NS and Krishna Raketla. Sentiment analysis of drug reviews using transfer learning. In *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 1794–1799, Coimbatore, India, 2021. IEEE.

[193] Nir Ofek, Lior Rokach, Cornelia Caragea, and John Yen. The importance of pronouns to sentiment analysis: Online cancer survivor network case study. In *WWW'15 Companion: Proceedings of the 24th International Conference on World Wide Web*, pages 83–84, Florence, Italy, 2015. ACM.

[194] Bahadorreza Ofoghi, Meghan Mann, and Karin Verspoor. Towards early discovery of salient health threats: A social media emotion classification technique. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, pages 504–515, Kohala Coast, Hawaii, USA, 2016. World Scientific.

[195] Atte Oksanen, David Garcia, Anu Sirola, Matti Näsi, Markus Kaakinen, Teo Keipi, and Pekka Räsänen. Pro-anorexia and anti-pro-anorexia videos on YouTube: Sentiment analysis of user responses. *Journal of Medical Internet Research*, 17(11):e256, 2015.

[196] Nels Oscar, Pamela A. Fox, Racheal Croucher, Riana Wernick, Jessica Keune, and Karen Hooker. Machine learning, sentiment analysis, and tweets: an examination of Alzheimer's disease stigma on Twitter. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 72(5):742–751, 2017.

[197] Youssef Oualil, Mittul Singh, Clayton Greenberg, and Dietrich Klakow. Long-short range context neural networks for language modeling. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1473–1481, Austin, Texas, 2016. Association for Computational Linguistics.

[198] Can Özbey, Berke Dilekoğlu, and Sevim Açiksöz. The impact of ensemble learning in sentiment analysis under domain shift. In *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6. IEEE, 2021.

[199] Rajesh R. Pai and Sreejith Alathur. Assessing mobile health applications with Twitter analytics. *International Journal of Medical Informatics*, 113:72–84, 2018.

[200] Alexander Pak, Delphine Bernhard, Patrick Paroubek, and Cyril Grouin. A combined approach to emotion detection in suicide notes. *Biomedical Informatics Insights*, 5(Suppl. 1):105–114, 2012.

[201] Marco Palomino, Tim Taylor, Ayse Göker, John Isaacs, and Sara Warber. The online dissemination of nature–health concepts: Lessons from sentiment analysis of social media relating to "nature-deficit disorder". *International Journal of Environmental Research and Public Health*, 13(1):142, 2016.

[202] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135, 2008.

[203] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS 2019)*, volume 32, pages 8024–8035, Vancouver, BC, Canada, 2019.

[204] Ted Pedersen. Rule-based and lightly supervised methods to predict emotions in suicide notes. *Biomedical Informatics Insights*, 5(Suppl. 1):185–193, 2012.

[205] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics.

[206] Martín Pérez-Pérez, Gael Pérez-Rodríguez, Florentino Fdez-Riverola, Anália Lourenço, et al. Using Twitter to understand the human bowel disease community: Exploratory analysis of key topics. *Journal of medical Internet research*, 21(8):e12610, 2019.

[207] John P. Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, K. Bretonnel Cohen, John Hurdle, and Christopher Brew. Sentiment analysis of suicide notes: A shared task. *Biomedical Informatics Insights*, 5(Suppl. 1):3–16, 2012.

[208] Kenneth Portier, Greta E. Greer, Lior Rokach, Nir Ofek, Yafei Wang, Prakhar Biyani, Mo Yu, Siddhartha Banerjee, Kang Zhao, Prasenjit Mitra, and John Yen. Understanding topics and sentiment in an online cancer survivor community. *Journal of the National Cancer Institute Monographs*, 2013(47):195–198, 2013.

[209] Qiao Qian, Minlie Huang, Jinhao Lei, and Xiaoyan Zhu. Linguistically regularized LSTM for sentiment classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1679–1689, Vancouver, Canada, 2017. Association for Computational Linguistics.

[210] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[211] Viju Raghupathi, Jie Ren, and Wullianallur Raghupathi. Studying public perception about vaccination: A sentiment analysis of tweets. *International journal of environmental research and public health*, 17(10):3464, 2020.

[212] Sreeram Ramagopalan, Radek Wasiak, and Andrew P. Cox. Using Twitter to investigate opinions about multiple sclerosis treatments: a descriptive, exploratory study. *F1000Research*, 3, 2014.

[213] Jyoti Ramteke, Samarth Shah, Darshan Godhia, and Aadil Shaikh. Election result prediction using Twitter sentiment analysis. In *2016 International Conference on Inventive Computation Technologies (ICICT 2016)*, pages 1–5, Coimbatore, India, 2016. IEEE.

[214] Majid Rastegar-Mojarad, Zhan Ye, Daniel Wall, Narayana Murali, and Simon Lin. Collecting and analyzing patient experiences of health care from social media. *JMIR Research Protocols*, 4(3):e3433, 2015.

[215] Vinit Ravishankar, Artur Kulmizev, Mostafa Abdou, Anders Søgaard, and Joakim Nivre. Attention can reflect syntactic structure (if you let it). In *16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (EACL)*, pages 3031–3045, Online, 2021. Association for Computational Linguistics.

[216] Jonathon Read, Erik Velldal, and Lilja Øvrelid. Labeling emotions in suicide notes: Cost-sensitive learning with heterogeneous features. *Biomedical Informatics Insights*, 5(Suppl. 1):99–103, 2012.

[217] Benjamin J. Ricard, Lisa A. Marsch, Benjamin Crosier, and Saeed Hassanpour. Exploring the utility of community-generated social media content for detecting depression: An analytical study on Instagram. *Journal of Medical Internet Research*, 20(12):e11817, 2018.

[218] Kirk Roberts and Sanda M. Harabagiu. Statistical and similarity methods for classifying emotion in suicide notes. *Biomedical Informatics Insights*, 5(Suppl. 1):195–204, 2012.

[219] Marco Roccetti, Gustavo Marfia, Paola Salomoni, Catia Prandi, Rocco Maurizio Zagari, Faustine Linda Gningaye Kengni, Franco Bazzoli, and Marco Montagnani. Attitudes of Crohn's disease patients: Infodemiology case study and sentiment analysis of Facebook and Twitter posts. *JMIR Public Health and Surveillance*, 3(3):e7004, 2017.

[220] Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. A hierarchical model of reviews for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 999–1005, Austin, TX, USA, 2016. Association for Computational Linguistics.

[221] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[222] María del Pilar Salas-Zárate, Jose Medina-Moreira, Katty Lagos-Ortiz, Harry Luna-Aveiga, Miguel Angel Rodriguez-Garcia, and Rafael Valencia-Garcia. Sentiment analysis on Tweets about diabetes: An aspect-level approach. *Computational and Mathematical Methods in Medicine*, 2017:5140631, 2017.

[223] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019.

[224] Karthik V. Sarma, Brennan M.R. Spiegel, Mark W. Reid, Shawn Chen, Raina M. Merchant, Emily Seltzer, and Corey W. Arnold. Estimating the health-related quality of life of Twitter users using semantic processing. *Studies in Health Technology and Informatics*, 264:1065–1069, 2019.

[225] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *Lecture Notes in Computer Science, 15th International Conference on European Semantic Web Conference (ESWC 2018)*, volume 10843, pages 593–607, Heraklion, Greece, 2018. Springer International Publishing.

[226] Kim Schouten and Flavius Frasincar. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830, 2016.

[227] Bonggun Shin, Timothy Lee, and Jinho D. Choi. Lexicon integrated CNN models with attention for sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–158, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[228] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic

compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, WA, USA, 2013.

[229] Sunghwan Sohn, Manabu Torii, Dingcheng Li, Kavishwar Wagholikar, Stephen Wu, and Hongfang Liu. A hybrid approach to sentiment sentence classification in suicide notes. *Biomedical Informatics Insights*, 5(Suppl. 1):43–50, 2012.

[230] Irena Spasić, Pete Burnap, Mark Greenwood, and Michael Arribas-Ayllon. A naïve Bayes approach to classifying topics in suicide notes. *Biomedical Informatics Insights*, 5(Suppl. 1):87–97, 2012.

[231] Irena Spasić and Goran Nenadić. Clinical text data in machine learning: Systematic review. *JMIR Medical Informatics*, 8(3):e17984, 2020.

[232] Irena Spasić, David Owen, Andrew Smith, and Kate Button. KLOSURE: Closing in on open–ended patient questionnaires with text mining. *Journal of Biomedical Semantics*, 10(1):1–11, 2019.

[233] Irena Spasić, Özlem Uzuner, and Li Zhou. Emerging clinical applications of text analytics. *International Journal of Medical Informatics*, 134:103974, 2020.

[234] Dominik Spinczyk, Mateusz Bas, Mariusz Dzieciątko, Michał Maćkowski, Katarzyna Rojewska, and Stella Maćkowska. Computer-aided therapeutic diagnosis for anorexia. *BioMedical Engineering OnLine*, 19:53:1–23, 2020.

[235] Philip J. Stone, Dexter C. Dunphy, and Marshall S. Smith. *The general inquirer: A Computer Approach to Content Analysis.* MIT press, 1966.

[236] Carlo Strapparava and Alessandro Valitutti. WordNet Affect: an affective extension of WordNet. In *LREC 2004, The Fourth International Conference on Language Resources and Evaluation*, pages 1083–1086, Lisbon, Portugal, 2004. European Language Resources Association (ELRA).

[237] Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 380–385, Minneapolis, MN, USA, 2019. Association for Computational Linguistics.

[238] Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. Aspect-level sentiment analysis via convolution over dependency tree. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5683–5692, Hong Kong, China, 2019. Association for Computational Linguistics.

[239] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML'17: Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 3319–3328, Sydney, Australia, 2017. JMLR.org.

[240] Maite Taboada, Caroline Anthony, and Kimberly D. Voll. Methods for creating semantic orientation dictionaries. In *LREC 2006, The Fifth International Conference on Language Resources and Evaluation*, pages 427–432, Genoa, Italy, 2006. European Language Resources Association (ELRA).

[241] Patrick J. Tighe, Ryan C. Goldsmith, Michael Gravenstein, H. Russell Bernard, and Roger B. Fillingim. The painful tweet: text, sentiment, and community structure analyses of tweets pertaining to pain. *Journal of Medical Internet Research*, 17(4):e84, 2015.

[242] Le-Thuy T. Tran, Guy Divita, Marjorie E. Carter, Joshua Judd, Matthew H. Samore, and Adi V. Gundlapalli. Exploiting the UMLS Metathesaurus for extracting

and categorizing concepts representing signs and symptoms to anatomically related organ systems. *Journal of Biomedical Informatics*, 58:19–27, 2015.

[243] Peter D. Turney and Michael L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346, 2003.

[244] Alessandro Valitutti, Carlo Strapparava, and Oliviero Stock. Developing affective lexical resources. *PsychNology J.*, 2(1):61–83, 2004.

[245] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS 2017)*, volume 30, pages 5998–6008, Long Beach, CA, USA, 2017. Curran Associates, Inc.

[246] Byron C. Wallace, Michael J. Paul, Urmimala Sarkar, Thomas A. Trikalinos, and Mark Dredze. A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *Journal of the American Medical Informatics Association*, 21(6):1098–1103, 2014.

[247] Tao Wang, Emmanouil Mentzakis, Markus Brede, and Antonella Ianni. Estimating determinants of attrition in eating disorder communities on Twitter: an Instrumental variables approach. *Journal of Medical Internet Research*, 21(5):e10942, 2019.

[248] Wenbo Wang, Lu Chen, Ming Tan, Shaojun Wang, and Amit P. Sheth. Discovering fine-grained sentiment in suicide notes. *Biomedical Informatics Insights*, 5(Suppl. 1):137–145, 2012.

[249] Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. Recursive neural conditional random fields for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 616–626, Austin, TX, USA, 2016. Association for Computational Linguistics.

[250] Xingyou Wang, Weijie Jiang, and Zhiyong Luo. Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2428–2437, Osaka, Japan, 2016. The COLING 2016 Organizing Committee.

[251] Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, TX, USA, 2016. Association for Computational Linguistics.

[252] Richard Wicentowski and Matthew R. Sydes. Emotion detection in suicide notes using maximum entropy classification. *Biomedical Informatics Insights*, 5(Suppl. 1):51–60, 2012.

[253] Janyce M. Wiebe and Rebecca F. Bruce. Probabilistic classifiers for tracking point of view. In *Proceedings of the 1995 AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 181–187, Menlo Park, CA, USA, 1995.

[254] Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 246–253, College Park, MD, USA, 1999. Association for Computational Linguistics.

[255] Matthew T. Wiley, Canghong Jin, Vagelis Hristidis, and Kevin M. Esterling. Pharmaceutical drugs chatter on online social networks. *Journal of Biomedical Informatics*, 49:245–254, 2014.

[256] Lowri Williams, Michael Arribas-Ayllon, Andreas Artemiou, and Irena Spasić. Comparing the utility of different classification schemes for emotive language analysis. *Journal of Classification*, 36(3):619–648, 2019.

[257] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language-Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, BC, Canada, 2005. Association for Computational Linguistics.

[258] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, RÃ©mi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceeding of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, pages 38–45, Online, 2020. Association for Computational Linguistics.

[259] David H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390, 1996.

[260] Chuhan Wu, Fangzhao Wu, Junxin Liu, Yongfeng Huang, and Xing Xie. Sentiment lexicon enhanced neural sentiment classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1091–1100, Beijing, China, 2019. ACM.

[261] Haiyan Wu, Zhiqiang Zhang, Shaoyun Shi, Qingfeng Wu, and Haiyu Song. Phrase dependency relational graph attention network for aspect-based sentiment analysis. *Knowledge-Based Systems*, 236:107736, 2022.

[262] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex

Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.

[263] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S. Yu Philip. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2020.

[264] Luwei Xiao, Yun Xue, Hua Wang, Xiaohui Hu, Donghong Gu, and Yongsheng Zhu. Exploring fine-grained syntactic information for aspect-based sentiment classification with dual graph neural networks. *Neurocomputing*, 471:48–59, 2022.

[265] Hu Xu, Bing Liu, Lei Shu, and Philip Yu. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2324–2335, Minneapolis, MN, USA, 2019. Association for Computational Linguistics.

[266] Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. Double embeddings and CNN-based sequence labeling for aspect extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 592–598, Melbourne, Australia, 2018. Association for Computational Linguistics.

[267] Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2324–2335, Minneapolis, MN, USA, 2019. Association for Computational Linguistic.

[268] Yan Xu, Yue Wang, Jiahua Liu, Zhuowen Tu, Jian-Tao Sun, Junichi Tsujii, and Eric Chang. Suicide note sentiment classification: a supervised approach augmented by web data. *Biomedical Informatics Insights*, 5(Suppl. 1):31–41, 2012.

[269] Shweta Yadav, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. Medical sentiment analysis using social media: Towards building a patient assisted system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan, May 2018. European Language Resources Association.

[270] Fu-Chen Yang, Anthony J.T. Lee, and Sz-Chen Kuo. Mining health social media with sentiment analysis. *Journal of Medical Systems*, 40(11):1–8, 2016.

[271] Hui Yang, Alistair Willis, Anne De Roeck, and Bashar Nuseibeh. A hybrid model for automatic emotion recognition in suicide notes. *Biomedical Informatics Insights*, 5(Suppl. 1):17–30, 2012.

[272] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems (NeurIPS 2019)*, volume 32, pages 5753–5763, Vancouver, BC, Canada, 2019. Curran Associates, Inc.

[273] Eric Yeh, William Jarrold, and Joshua Jordan. Leveraging psycholinguistic resources and emotional sequence models for suicide note emotion annotation. *Biomedical Informatics Insights*, 5(Suppl. 1):155–163, 2012.

[274] Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. Comparative study of CNN and RNN for natural language processing. *CoRR*, abs/1702.01923, 2017.

[275] Sunmoo Yoon and Suzanne Bakken. Methods of knowledge discovery in tweets. In *NI 2012: 11th International Congress on Nursing Informatics*, volume

2012:463, Montreal, Qc, Canada, 2012. American Medical Informatics Association.

[276] Ning Yu, Sandra Kübler, Joshua Herring, Yu-Yin Hsu, Ross Israel, and Charese Smiley. LASSA: emotion detection via information fusion. *Biomedical Informatics Insights*, 5(Suppl. 1):71–76, 2012.

[277] Li Zhang, Haimeng Fan, Chengxia Peng, Guozheng Rao, and Qing Cong. Sentiment analysis methods for HPV vaccines related tweets based on transfer learning. *Healthcare*, 8(3):307, 2020.

[278] Ling Zhang, Magie Hall, and Dhundy Bastola. Utilizing Twitter data for analysis of chemotherapy. *International Journal of Medical Informatics*, 120:92–100, 2018.

[279] Shaodian Zhang, Erin Bantum, Jason Owen, and Noémie Elhadad. Does sustained participation in an online health community affect sentiment? In *AMIA Annual Symposium Proceedings*, volume 2014, pages 1970–1979, Washington, DC, USA, 2014. American Medical Informatics Association.

[280] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems*, volume 1, pages 649–657, Montreal, Canada, 2015. MIT Press.

[281] Yuhao Zhang, Peng Qi, and Christopher D. Manning. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, Brussels, Belgium, 2018. Association for Computational Linguistics.

[282] Kang Zhao, John Yen, Greta Greer, Baojun Qiu, Prasenjit Mitra, and Kenneth Portier. Finding influential users of online health communities: a new metric

based on sentiment influence. *Journal of the American Medical Informatics Association*, 21(e2):e212–e218, 2014.

[283] Meng Zhao, Jing Yang, Jianpei Zhang, and Shenglong Wang. Aggregated graph convolutional networks for aspect-based sentiment classification. *Information Sciences*, 600:73–93, 2022.

[284] Xujuan Zhou, Enrico Coiera, Guy Tsafnat, Diana Arachi, Mei-Sing Ong, and Adam G. Dunn. Using social connection information to improve opinion mining: Identifying negative sentiment about HPV vaccines on Twitter. *Studies in Health Technology and Informatics*, 216:761–765, 2015.

[285] Anastazia Žunić, Padraig Corcoran, and Irena Spasić. Improving the performance of sentiment analysis in health and wellbeing using domain knowledge. In *HealTAC 2020: Healthcare Text Analytics Conference*, London, UK (Online), 2020.

[286] Anastazia Žunić, Padraig Corcoran, and Irena Spasić. Sentiment analysis in health and well-being: Systematic review. *JMIR Medical Informatics*, 8(1):e16023, 2020.

[287] Anastazia Žunić, Padraig Corcoran, and Irena Spasić. Aspect-based sentiment analysis with graph convolution over syntactic dependencies. *Artificial Intelligence in Medicine*, 119:102138, 2021.

[288] Anastazia Žunić, Padraig Corcoran, and Irena Spasić. The case of aspect in sentiment analysis: Seeking attention or co-dependency? *Machine Learning and Knowledge Extraction*, 4(2):474–487, 2022.