REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mohamed Khider – BISKRA

Faculté des Sciences Exactes, des Sciences de la Nature et de la Vie

Département d'informatique

Laboratoire de L'Informatique Intelligente (Linfi)

N° Série :…………

THESE

Présentée pour obtenir le grade de

DOCTORAT EN LMD EN INFORMATIQUE

Par

Melle NAHILI Wedjdane

THEME

Une approche connexionniste pour l'intelligence territoriale basée sur les réseaux sociaux

Soutenue le : 22/10/2020
Devant le jury composé de :

KAZAR Okba	**Professeur à l'Université de Biskra**	**Président**
REZEG Khaled	**M.C.A à l'Université de Biskra**	**Rapporteur**
MELKEMI Kamal-Eddine	**Professeur à l'Université de Batna 2**	**Examinateur**
SEGHIR Rachid	**Professeur à l'Université de Batna 2**	**Examinateur**
KAHLOUL Laid	**Professeur à l'Université de Biskra**	**Examinateur**

2019/2020

'Those who think they can and those who think they can't are both usually right'

- Confucius

Acknowledgements

This dissertation got completed with the guidance and help of several individuals, who in one way or another contributed and extended their valuable support in the preparation and completion of this research…,

My first and foremost gratitude to my Supervisor Doctor Rezeg Khaled Senior Lecturer at Biskra University. Whose guidance and encouragement is unforgettable. I sincerely pay my gratitude for his consistent encouragement along the way. The suggestions for improvement in work and review feedback have given me knowledge, to step ahead to the next level of study…,

I would like to thank the members of my dissertation committee. First, the president of the committee Professor Kazar Okba from Biskra University. His availability and sincerity, and also his constant support were remarkable. His gentile way of pushing me to bring the best in me. Second, a distinct thank you goes to Professor Melkemi Kamal-Eddine and Professor Seghir Rachid from Batna 2 University, also Professor Kahloul Laid from Biskra University - not only for their time and extreme patience, but for the interest they have shown in my work by taking place in this committee. Their insightful discussion and intellectual contributions to my development as a researcher…

I would also like to thank my parents, my source of strength and pride, I pay them my deepest gratitude, love and respect. Who have always supported and encouraged me in every walk…

A noteworthy gratitude goes out to my source of light in this life, this accomplishment would not have been possible without you Mehdi. Thank you for everything…

I would like to acknowledge my second parents for their constant support. They believed in me, in all my endeavours and who so lovingly and unselfishly, cared for me…

Finally, a special thank you to my brothers from another mother, my source of laughter no matter the circumstances…

Abstract

Long before the invention of the Internet, the purchasing process and customer behaviour were supported by the word-of-mouth, as it was the only channel to acquire feedback and customer reviews. In many cases, our buying choices were made with a leap of faith and hope that our purchase turned out to be everything we expected. But with the rise of Web 2.0, customers share information or opinions about products and services, politics, current events online. As result, people and organisations refer to these information to harvest valuable insights and hence, make intelligent decisions. This shared information is a gold mine, if leveraged effectively, can provide rich and valuable insights. The problem with this information is that it is informal and unstructured, thus, difficult to assess automatically and in huge volume. Accordingly, these data require appropriate processing to obtain useful information. Sentiment analysis (SA) is used to extract knowledge from online data. Research in the field of SA seek to extract sentiment from textual data. In this thesis, two approaches are provided to conduct sentiment analysis on text. The first one is a lexicon-based approach for multi-class Twitter sentiment analysis by developing a sentiment lexicon specific to the social media domain. The second one is a deep learning approach for binary-class sentiment analysis of reviews by proposing a convolutional neural network (CNN). This research uses universally accessible data, i.e Twitter and movie reviews datasets to evaluate the proposed frameworks for their reliability and validity. Experiments were conducted using the proposed methodologies; firstly, the lexicon-based approach was evaluated on Twitter data. The results show that the developed lexicon is able to capture sentiment intensity and handle social media text. Secondly, the proposed CNN model was trained and tested using the IMDb dataset. For evaluation, accuracy was used. A sizeable performance improvement was reported whereby the proposed network yielded better results compared to prior models from the related work.

Keywords: sentiment analysis, natural language processing, text analytics, text mining, deep learning, convolutional neural networks, IMDb dataset

Resumé

Bien avant l'invention de l'Internet, le processus d'achat et le comportement des clients était soutenus par le bouche-à-oreille, étant le seul canal pour acquérir le feedback des clients. Auparavant, nos choix d'achat était faits sur un acte de foi, en espérant que cet achat allait s'avérer être tout ce que nous attendions. Mais avec l'essor du Web 2.0, les clients partagent des informations sur les produits et les services en ligne. En conséquence, les utilisateurs et les organisations utilisent ces données pour récolter des informations précieuses afin de rendre leurs décisions intelligentes. Ces données sont une mine d'or, si elles sont utilisées efficacement, elles peuvent fournir des informations précieuses. Néanmoins, le problème avec ces données c'est qu'elles sont informelles et non structurées, donc difficiles à traiter automatiquement et en volume énorme. Ainsi, l'analyse des sentiments (AS) est utilisée pour extraire les connaissances des données. L'AS consiste à prélever le sentiment exprimé dans un texte. Dans cette thèse, deux approches sont proposées pour prédire le sentiment de données en ligne. La première est une approche sémantique pour identifier le sentiment énoncé dans des tweets. Pour cela, un dictionnaire a été développé. La seconde est une approche deep learning basée sur un réseau de neurones convolutionnels conçu pour distinguer le sentiment des critiques de films. Dans ce travail, des données publiquement accessibles sont utilisées afin d'évaluer les approches proposées en matière de fiabilité et validité. Des expérimentations ont été menées en employant les approches proposées; premièrement, l'approche sémantique a été évaluée sur des données Twitter. Les résultats montrent que le dictionnaire est non seulement apte d'extraire le sentiment mais aussi de capturer son l'intensité. Deuxièmement, la base de données IMDb a été utilisée pour l'apprentissage du modèle proposé. Pour l'évaluation, la métrique de précision a été utilisée. Le bilan démontre une amélioration considérable en termes de performance, où le modèle proposé a obtenu de meilleurs résultats par rapport aux modèles de travaux relatifs.

Mots clés: analyse de sentiments, traitement langage naturel, text mining, apprentissage profond, réseau de neurones convolutionnels, IMDb dataset

ملخص

قبل اختراع الإنترنت بفترة طويلة ، كانت عملية الشراء و سلوك العملاء مدعوما من خلال الحديث الشفوي، حيث كانت القناة الوحيدة التي تحصل على رجع صدى ومراجعات العملاء. في كثير من الحالات، تم إجراء اختياراتنا للشراء بقفزة من الإيمان و نأمل أن تكون عملية الشراء الخاصة بنا أن تكون كل ما كنا نتوقع. ولكن مع ظهور الويب ٢.٠، يشارك العملاء المعلومات أو الآراء حول المنتجات والخدمات والسياسة والأحداث الجارية على المباشر. ونتيجة لذلك، يشير الأشخاص والمنظمات إلى هذه المعلومات لجني على رؤى قيمة، وبالتالي اتخاذ قرارات ذكية. هذه المعلومات المشتركة هي كنز من ذهب، يمكن أن تقدم رؤى غنية وقيمة إذا تم الاستفادة منها بشكل فعال. تكمن المشكلة في هذه المعلومات في أنها عامية وغير منظمة، وبالتالي، من الصعب تقييمها تلقائيا و بكميات ضخمة، وبناءً على ذلك، تتطلب هذه البيانات معالجة مناسبة لكسب معلومات مفيدة. يستخدم تحليل المشاعر لاستخراج المعرفة من بيانات الخط. البحث في مجال SA يسعى لاستخراج التوجيه من البيانات النصية في هذه الأطروحة، تم توفير طريقتين لإجراء تحليل التوجه على النص: الأول هو نهج قائم على المعجم لتحليل توجه من أجل تعدد الفئات Twitter من خلال تطوير توجه خاص بمجال الوسائط الاجتماعية.والثاني هو نهج التعلم العميق لتحليل توجهات الطبقة الثنائية للمراجعات من خلال اقتراح شبكة عصبية تلافيفية CNN. يستخدم هذا البحث بيانات متاحة عالميا،مجموعات بيانات Twitter و عروض أفلام لتقييم الأطر المقترحة لموثوقيتها و صلاحيتها. أجريت التجارب باستخدام المنهجيات المقترحة: أولاً، تم تقييم النهج القائم على المعجم على بيانات. أظهرت النتائج أن المعجم المتطور قادر على التقاط شدة التوجه والتعامل مع الطب الاجتماعي. ثانياًا ، تم تدريب واختبار نموذج CNN المقترح باستخدام مجموعة بيانات IMDb؛ و للتقييم، تم استخدام الدقة. تم الإعلان عن تحسن كبير في الأداء حيث حققت الشبكة المقترحة نتائج أفضل مقارنة بالنماذج الأولية من العمل المتعلق.

الكلمات المفتاحية: تحليل المشاعر، معالجة اللغة الطبيعية، تحليلات النص، تحليل النصوص، تعلم عميق، الشبكات العصبية التلافيفية، مجموعة البيانات IMDb

Contents

List of Figures

List of Tables

Abbreviations

SA	**S**entiment **A**nalysis
AS	**A**nalyse de **S**entiments
NLP	**N**atural **L**anguage **P**rocessing
UGC	**U**ser **G**enerated **C**ontent
CNN	**C**onvolutional **N**eural **N**etwork
emb-CNN	**emb**edding **C**onvolutional **N**eural **N**etwork
RNN	**R**eccurent **N**eural **N**etwork
NN	**N**eural **N**etwork
LSTM	**L**ong **S**hort **T**erm **M**emory
SVM	**S**upport **V**ector **M**achine
ME	**M**aximum **E**ntropy
LR	**L**ogistic **R**egression
NB	**N**aive **B**ayes
BNB	**B**enoulli **N**aive **B**ayes
MNB	**M**ultinomial **N**aive **B**ayes
SGD	**S**tochastic **G**radient **D**escent
POS	**P**art **O**f **S**peech
BoW	**B**ag **of** **W**ords
SL	**S**upervised **L**earning
LB	**L**exicon **B**ased
DB	**D**ictionary**B**ased
TF-IDF	**T**erm **F**requency textbfInverse **D**ocument **F**requency

Chapter One

Introduction

1.1 Research background

The spectacular diffusion of information and communication technologies (ICT) has resulted in an exponential increase in the information available to socio-economic actors. As stated by (Pateyron, 1998) more information has already been produced during the last thirty years than in the previous ten thousand years, and it is expected to double every five years. As result, it has become vital to analyse the generated content by information flows from different sources such as social networks. The new challenge does not lie on how to access information but how to select and extract the relevant information to make good strategic decisions rapidly when surrounded by a competitive environment (Pelissier and Pybourdin, 2009). In this perspective, the ability of acquiring and analysing these data by actors such as organisations is becoming an essential skill to achieve success. In this context territorial intelligence, originally defined as *"all coordinated actions of research, processing and distribution in view of exploiting useful information for economic actors"* (Matre, 1994) appeared as an approach responding to the challenges faced during decision-making.

In real life, facts are certainly vital for individuals or businesses to make judgements or decisions, but opinions or beliefs also play essential roles. Human behaviour is considerably affected by their subjective feelings, such as attitude, emotion, opinion or sentiment. The decisions we take can be influenced by others' point of views or perceptions of the world to a considerable degree, because carrying others' opinions is wired into all human beings naturally. The way people value things can easily be affected by others' sentiments towards them, taking this aspect into consideration has become an important factor in decision-making processes (Campbell et al., 2010).

Therefore, it is crucial to mine people's opinions and feelings about a certain subject matter of interest, which is the task of sentiment analysis (Dave et al., 2003, Esuli and Sebastiani, 2006, Pang and Lee, 2008). More specifically, sentiment analysis also called *opinion mining* is a sub-field of natural language processing (NLP) which is a type of text classification that primarily deals with subjective text, aiming to extract and classify sentiments or opinions present in natural language text by using computational methods (Pang and Lee, 2008, Liu, 2015).

With the increasing quality of the Web and social media sites such as Facebook, Twitter and Instagram, the studies of sentiment analysis have drawn a lot of attention from researchers due to the fact that Internet access has provided a simple and effective platform for individuals (users) to share opinions concerning everything. Before the emergence of World Wide Web [1], many people typically asked friends or family for advice or recommendations before buying products. The evolution of Web 2.0 [2] has provided individuals with the chance to post their thoughts and opinions covering a wide variety of topics on different platforms. As John Scalzit states: *"Everyone is entitled to their opinion about the things they read, or watch, or listen to, or taste, or whatever. They're also entitled to express them online"*. With the growth of the digital age, the number of people able to share their thoughts online and also be aware of others' opinions is increasing. Thus, there is a tremendous amount of data containing opinions generated from different sources such as reviews, posts from blogs, micro-blogs like Twitter or other social media [3] (George, 2015). These data are also referred to as user-generated content (UGC), which represents any positive or negative communication created by costumers regarding a product or company, which is openly shared and made available via the Internet.

Compared with the traditional strategies of promotion in mass media, like televisions and newspapers, the online UGC is judged to be more reliable and balanced than those provided by businesses (Mudambi and Schuff, 2010, Chong et al., 2016). Thus, several organisations have integrated social media as part of their marketing/promoting strategies and encourage online users to share positive comments about their products and services on purpose. Online reviews, as a specific form of UGC, have

[1] The World Wide Web is a global information medium connected to the Internet, which enables users to search for information from one document to another (Berners et al., 2000).

[2] Web 2.0 are World Wide Web sites that highlight user-generated content, where users can read, write and share various content (O'Reilly, 2005).

[3] Social media also refers to a social networking service, which is an online platform that allows users to create user-generated content to share information, ideas and personal messages etc., in which, users can also develop social relations with others. It contains many different forms including blogs, forums, social gaming and so on (Obar and Wildman, 2015)

not solely become a major data source to assist customers make decisions, but also represent the base on which consumers re-evaluate their buying decisions and ultimately modify their shopping behaviour (Ye et al., 2011, Cantallops and Salvi, 2014), in line with various studies:

- According to (Shrestha, 2016), 92% of consumers read online product reviews before purchasing the products compared to 88% in 2014 and 40% of consumers make their initial decision by reading just one to three reviews;
- A study conducted by Nielsen shows that 66% of people trust online product reviews (Stone, 2015);
- A study indicates that between one to three negative online reviews could be enough to decrease 67% number of customers (Charlton, 2012);
- According to (Fan and Fuel, 2016), 97% of online customers read reviews before making any buying decisions, 32% say written reviews are the only element that makes them believe a site's reviews are relevant or useful, 94% typically read written reviews, 73% say written reviews are more taking into consideration than star or number ratings and 35% say that it takes only one negative review to make them decide not to buy a product.

The data statistics above uncover the interest that customers convey in online reviews about items, products and services as far as depending on the online suggestions or recommendations to settle on buying decisions. Therefore, the demand for sentiment analysis is critical as a result of this rush of interest. For customers, it is helpful for producing intelligent choices (decisions) by knowing the products' positive and negative attributes. As for companies or businesses, on the opposite hand, they can use it to harvest valuable insights about how customers really feel about them and improve their marketing methods and differentiate themselves from competitors so as to enhance their products and the customer experience (Wallace, 2015, Jain and Kumar, 2016). Moreover, understanding the opinions expressed in social media websites offers the possibility to bring tremendous business opportunities and great assistance in the decision making process (Alghamdi, 2013). With the aim to meet the need of customers and companies, effective and efficient sentiment analysis systems are requested in order to extract opinions and yield clear insights from online UGC.

1.2 Problem statement and research questions

The explosion of the Web 2.0 has not solely brought us researchers a huge volume of unstructured data represented in digital form, but also offered us great opportunity to interpret the sentiment of the public by analysing these large-scale data. However, all of the user-generated content is a double-edged sword: the larger the size of the data, the trickier it is to extract useful information. A survey shows that Facebook generates 250 million posts per hour and Twitter users on the opposite hand generate 21 million tweets per hour (George, 2015). In 2018, the review website TripAdvisor [4] generated approximately 730 million reviews and opinions from users around the world (Statista, 2019). Facing such big data, studies have already disclosed that over half of online customers experience frustrations throughout their online shopping, therefore, they are not able to make an informed decision.

Although we are in the era of Web 2.0, submerged with large amounts of data every day, companies and organisations additionally face issues in analysing the unstructured data effectively because of the fact that this process is time consuming and resource demanding. A survey shows that three quarters of 2,100 organisations do not have a straightforward idea of what their customers think about them and nearly 31% of them find it difficult to measure customers' opinions (Michael, 2012). It is obvious that they do not lack the data sources of customers' opinions, but the overwhelming size of data and the complexity of addressing the subjectivity aspect, makes it difficult to extract helpful information for organisations.

The necessity to analyse these unstructured data naturally resulted in the rise of research in the field of sentiment analysis. Sentiment analysis has been one of the most active research areas in natural language processing (NLP) since 2002 (see Chapter 2). The main task of sentiment analysis is to automatically determine the sentiment orientation in a given document (Pang and Lee, 2008, Turney, 2002) and indicate its polarity whether it is positive, negative or neutral (Liu, 2010).

Currently research on sentiment analysis has been dominated by two approaches: first, the supervised learning approaches, which aim to build and train classifiers by choosing the appropriate features (see Chapter 2). Second, the semantic orientation

[4] TripAdvisor is worldwide travel website that was an early adopter of UGC, which provides reviews of travel-related content such as hotel reviews and restaurants reviews (TripAdvisor, 2020)

approaches, that involve calculating the overall polarity based on sentiment orientation of terms (words), phrases or documents(see Chapter 2). Since the latter approaches utilize lexical resources like lists of opinion words, lexicons and dictionaries, they are also known as lexicon-based approaches (Peng et al., 2003, Ding et al., 2008, Taboada et al., 2011). Hence in this thesis, the terms 'semantic orientation approach' and 'lexicon-based approach' are used interchangeably.

Many sentiment analysis tools and applications have been developed to mine the sentiments present in user generated content on the Web. However, the reported performances were found to be very poor due to the complexity of the task of sentiment analysis in addition to the usage of natural language (Mohammad et al., 2013, Ouyang et al., 2015, Maynard and Bontcheva, 2016, Houshmand, 2017). Basically, sentiment analysis is primarily a problem of natural language processing (NLP) which deals with unstructured data (Liu, 2012). Even thought various approaches have been proposed to implement sentiment analysis, it is still difficult to capture and interpret some linguistic features, such as negation and mix-opinion text. This results in low accuracy of sentiment classification (Vinodhini and Chandrasekaran, 2012, Park, 2015a, Khan et al., 2016). Due to the existing real-world challenges in dealing with big data and current research gaps (see Chapter 2 for more details), the research presented in this thesis is motivated to address the following research questions:

1. *How can online text data be automatically and accurately classified with respect to their sentiments?*
2. *How can the intensity of sentiment in online text data be effectively captured?*
3. *How can deep learning enhance the process of Data Analytics for companies so they can improve their decision making?*

The first research question covers the need to manage the massive quantity of online reviews in an automated manner and improve the performance of sentiment classification. The second research question emphasizes the necessity to capture the intensity of the opinions present in a given piece of text in order to reflect real life complex scenarios. The last one pursues to propose an enhanced supervised learning model for text analytics and provide valuable insight to businesses in order to improve their products or services.

1.3 Aim and objectives

The aim of this thesis is to explore an effective way to conduct fine-grained sentiment analysis by upgrading the performance of sentiment classification. To meet the need of this aim, there are three goals that this research has tried to reach. The first objective intends to handle the text that contains both positive and negative orientated opinions, because the majority of available data shows that positive and negative sentiments co-occur in the same document or sentence. Most documents or sentences will have both positive and negative views. Besides the intensity of sentiment can vary from one review to another.

Secondly, following the lexicon orientation approach (see Chapter 2) for sentiment analysis, a sentiment lexicon for analysing social media text needs to be constructed and, is employed to determine the polarity of a sentence. The sentiment lexicon contains words with their sentiment orientations. Since there is a variety of domains, words could be used differently and show opposite sentiment orientations in each domain (see Chapter 3). In addition to the fact that online text specifically product reviews include a diversity of abbreviations and slang language. These aspects are taking into consideration, thus the sentiment lexicon used for sentiment analysis is the key to obtaining more accurate results.

Furthermore, following the supervised learning approach for sentiment analysis (see Chapter 2), a framework designed to improve text analytics is proposed which consists of a new convolutional neural network model with an embedding layer (see Chapter 3). To predict the sentiment orientation of text in our case reviews, the proposed model is trained using the movie reviews dataset IMDb. At the same time, we implemented different features and used different configurations to build the proposed network (see Chapter 4). We develop our evaluation metrics based on accuracy to quantitatively compare the effectiveness of our model against prior models from the related work. Achieving these objectives should lead to two coherent sentiment analysis frameworks that are proposed in this research (see Chapter 3), which aims to improve the performance of sentiment classification and provide in-depth text analytics.

1.4 Structure of the Thesis

This thesis contains five chapters, which are organised following the phases illustrated in Figure 1.1. These phases incorporate: problem identification, objective definition, design and development, implementation and evaluation. The first phase, problem identification, is introduced in chapter one. The chapter of literature review also simplifies the identification of the problem, and then motivates this research. The second phase of objective definition is also described in chapter one in order to acquire the aim of this research. Following this is the third phase, design and development, in which both sentiment analysis frameworks are presented. The fourth phase is implementation which aims to convey how the proposed systems perform social media sentiment analysis, which is also presented in chapter four. Afterwards, the evaluation of the proposed sentiment analysis frameworks is done in this phase.

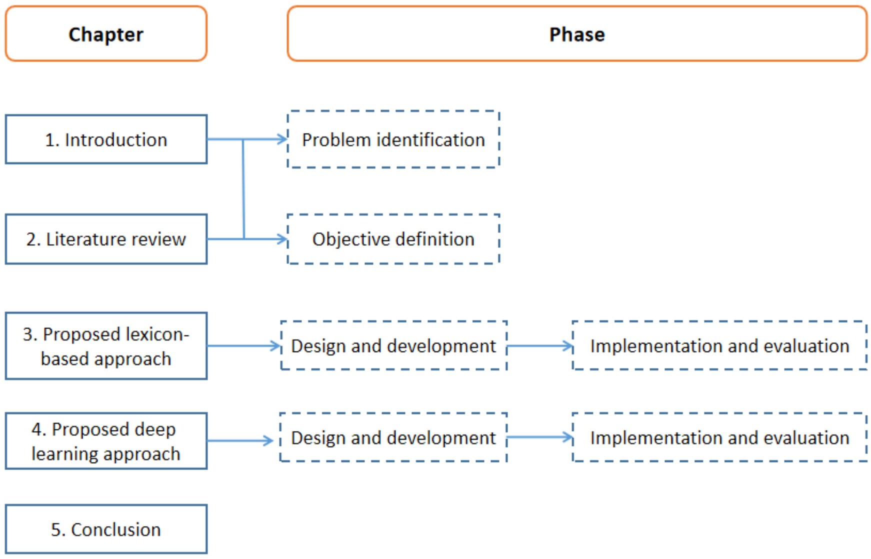


FIGURE 1.1: Thesis outline.

In chapter one, the scope of this research is presented with the introduction of the research background. The problems are also pinpointed, leading to the research questions in sentiment analysis, and the thesis aim and objectives that this research seeks to realise.

In chapter two, previous research on sentiment analysis are presented in detail. The chapter starts by introducing the notion of sentiment analysis and its different levels. In order to give an in-depth understanding of work on sentiment analysis, different progressive stages of sentiment analysis are also briefly presented. Furthermore, different approaches for sentiment analysis are surveyed in detail, and various literature gaps are identified, by which this thesis is motivated.

In chapter three, a lexicon-based approach for sentiment analysis of Twitter data is proposed. A description of the later approach is presented in detail to achieve the aim of this research. A system called TweetEcho is implemented following the proposed approach. Additionally, a sentiment lexicon specific to social media domain is developed. In order to show the output of our system, keywords like 'Samsung S10' and 'iphone X' are used for demonstration.

In chapter four, a deep learning approach for sentiment analysis using a new convolutional neural network with an embedding layer to classify movie reviews is introduced and implemented. Real data is used for training, testing and validating the proposed model. Different features are implemented and accuracy is used for evaluation. The experimental results are presented along with discussions.

Chapter five presents a summary of the work by reintroducing the research questions stated at the beginning of this thesis and explaining how each question has been addressed. Additionally, the contributions of this thesis are presented. Lastly, the limitations and future research directions are also proposed.

Chapter Two

Literature Review

2.1 Introduction

An important part of information-gathering is to determine what other people think about certain things. (Pang and Lee, 2008) state that new opportunities and challenges arise when trying to seek out and understand opinions of others due to growing availability of opinions sources such as online review sites and personal blogs. Lots of people do online research on a product, hotel or restaurant, before buying. The interest that individual users show in online opinions about products or services, and the potential influence of such opinions, is something that sellers of these items are paying more and more attention to. This is one of the reasons why sentiment analysis is an important task.

Since sentiment analysis in social media needs good knowledge of sentiment analysis techniques, this chapter will lay out a landscape or a detailed summary and a comprehensive understanding of the studies in the field of sentiment analysis, by beginning with introducing the notion of sentiment analysis. Since the studies of sentiment analysis have been applied in different levels, each level of sentiment analysis is investigated separately in this chapter, that is, document level, sentence level, aspect level and user level sentiment analysis.

Although the quick development of sentiment analysis corresponds with the explosion of Web 2.0, previous research related to sentiment analysis have set the foundation for the current research activity. For the purpose of having an in-depth understanding of the background and methods of sentiment analysis, various progressive stages of sentiment analysis are presented. Furthermore, two main approaches for sentiment analysis, which are the semantic orientation approach and the machine learning approach, will be reviewed in detail. Each approach will also be inspected and a comparison of both methods will be given in order to point out the advantages and drawbacks of each approach. At last, the research gaps will be discussed.

2.2 The impact of social networks

One of the defining phenomena of the present days reshaping the world as we know it, is the worldwide accessibility to the Internet. The major result of the World Wide Web is social media, which comes in many forms, including blogs, forums, business networks, photo-sharing platforms, social gaming, microblogs, chat apps, and last but not least social networks. Despite the universality of social networks, market potential is massively growing considering that on average, global internet users spend an enormous amount of time per day using social networks. This phenomenon bends brands and corporations on adapting their marketing strategies to take advantage of that time and screen usage to promote their products and services via social media marketing and advertising.

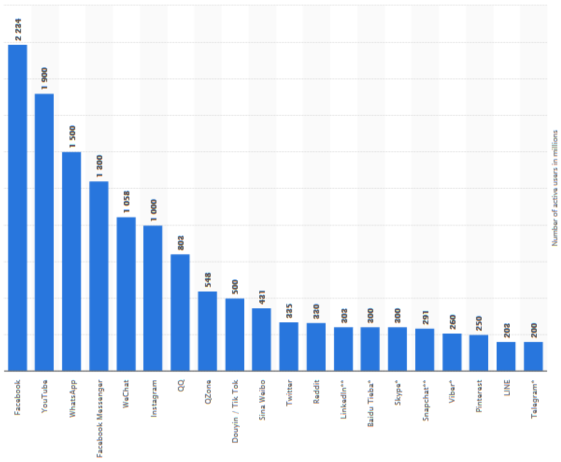


FIGURE 2.1: This statistic provides information on the most popular networks worldwide in October 2018 from (Statista, 2018).

As stated by (Kaplan and Haenlein, 2012), social media can be defined as the *'group of internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content'*. Over the past decade, besides the leaders of the World Wide Web such as Facebook, Twitter, YouTube, Google and LinkedIn, there are currently emerging market opportunities for a variety of other services which target different groups of users depending on the targeted niche and the particular data need. This enlarges the scope to very different kinds of research from consumer preferences and recommendation systems to decision making. Hence from the above Figure 2.1 and Figure 2.2, these statistics provide information on the most popular networks worldwide as of October 2018, classified by number of active accounts. According to (Statista, 2018), Facebook is the market leader with more than one billion registered accounts and currently has 2.23 billion monthly active users. These numbers show the number of social media users worldwide from 2010 to 2016 with predictions to 2021 (Statista, 2018).

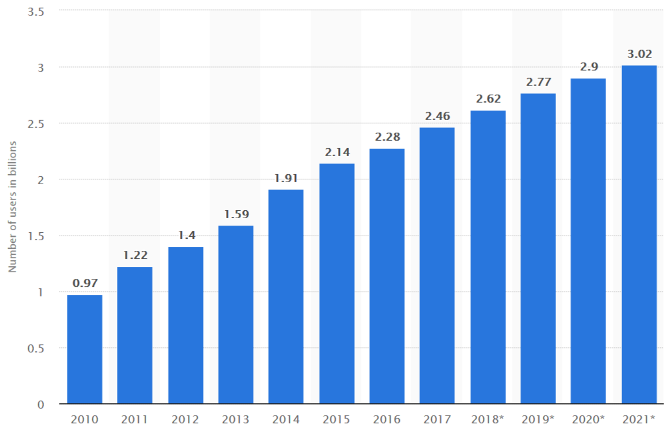


FIGURE 2.2: Number of active users in billions (Statista, 2018).

2.3 Sentiment Analysis in Social Media

Sentiment analysis is a sub-field of natural language processing (NLP) that aims to identify the sentiments or opinions, attitudes present in text or their sentiments towards specific topics. Sentiment describes a sentiment, opinion or an attitude expressed by an individual about an entity. Attitudes handle things like beliefs, preferences, and perceptions towards objects or persons (liking, loving, hating) (Scherer, 2003) are different from emotions (angry, sad, joyful, fearful, ashamed, proud) as result to external influences (Scherer, 2003). This differentiates sentiment analysis from other problems such as emotion analysis where the global emotional state is determined, and not the attitude towards a specific target. The strength and direction of sentiment (i.e., how positive or negative it is) is defined as its polarity. The simplest and most common polarity calculation supposes two categories, positive and negative. This assumption includes most voting systems used in research, such as:

- thumbs up/down (e.g., Facebook, YouTube, Netflix)
- positive, neutral, negative (e.g., eBay)
- star ratings (e.g., Amazon, IMDb, Rottentomatoes, Yelp)

Frequently, the sentiment polarity is represented inside the [-1; 1] interval, assuming that -1 represents the most negative polarity possible, and 1 represents the most positive one. There is some ambiguity concerning the center of the scale (0), which is commonly represented as neutral. It has been acknowledged that the neutral state is difficult to identify even for humans (Kim and Hovy, 2004), which is why it is sometimes excluded from experiments to simplify the problem (Speriosu et al., 2011, Go et al., 2009, da Silva et al., 2014, Saif et al., 2012, Blitzer et al., 2007, Bakliwal et al., 2012).

2.4 The notion of sentiment analysis

The word 'sentiment' is defined as a specific view or notion: 'opinion' and 'emotion' in Merriam Webster dictionary. As (Liu, 2010) expresses it "opinions are usually subjective expressions that describe people's sentiments, appraisals or feelings toward entities, events and their properties". Despite the fact that the field of sentiment analysis and opinion mining has recently attracted a lot of attention from researchers and marketers, there has been a steady interest of analysing opinions. Much of the early research on textual information analysis has been mainly focused on mining and retrieval of factual information, such as information retrieval, text classification or text clustering (Gupta and Lehal, 2009). One of the principal reasons for the lack of the research on sentiment analysis is that there was hardly no availability of opinionated text before the era of World Wide Web. As (Liu, 2010) describes that an individual often asks his/her friends or family for opinions before making any buying decision and an organisation normally conducted opinion votes and surveys to determine the users' sentiments towards its products or services. After the Internet started to be widely used due to the evolution of technology, people are able to express their opinions and emotions by posting reviews of products or services online as Liu deduces:

'This online word-of-mouth behaviour represents new and measurable sources of information with many practical applications'. Besides, (Pang and Lee, 2008) also pin point other factors which led to a huge blow up of research on sentiment analysis:*1) the rise of machine learning methods in natural language processing and information retrieval; 2) the availability of datasets for machine learning algorithms to be trained on, due to the growing of the World Wide Web websites; 3) the realisation of the captivating research opportunities and challenges and intelligence applications that the area provides'.* Particularly with the explosion of Web 2.0 platforms, such as blogs, Twitter, Facebook, Amazon and many other types of social media, any individual has direct access and ability to share his/her opinions and brand experience regarding any given product or service. Furthermore, organisations can alter their marketing strategies using social media monitoring and analysis. Nevertheless, it still can be a tricky task to obtain opinions sources and monitor them on the World Wide Web, because there is a large number of different sources. Also each source may have an enormous volume of user-generated content conveying sentiments or opinions. At the same time, another problem has manifested in trying to entitle this new area as (Zabin and Jefferies, 2008) described in their research report when they attempted to generate the consumer insights from online conversation:

'. . . the beginning of wisdom is the definition of terms, wrote Socrates. The aphorism is highly applicable when it comes to the world of social media monitoring and analysis, where any semblance of universal agreement on terminology is altogether lacking. Today, vendors, practitioners, and the media alike call this still-nascent arena everything from 'brand monitoring', 'buzz monitoring' and 'online anthropology', to 'market influence analytics', 'conversation mining' and 'online consumer intelligence'... The terms 'social media monitoring and analysis' best describes the content focus of this report. 'Social media' explains what is being monitored. 'Analysis' speaks to the fact that the process involves not just unleashing spiders that can crawl the Web and collect data in a mechanized fashion using natural text processing and other data mining and analytic technologies. The process also means making sense of the data, often with the help of human who can interpret and contextualise it in ways that machines aren't yet able to do, to generate actionable insights that results in smarter business decisions.'

The above quotation shows that it is essential to define a uniform terminology in the field of analysing consumers' online conversations (Pang and Lee, 2008). As a matter of fact, it is difficult to collect the appropriate sources, extract the information from the texts with sentiments and summarize them. Therefore, a framework is needed to automatically identify and analyse the online opinionated texts (Liu, 2010). In natural language processing (NLP), sentiment analysis covers various aspects regarding how information about sentiments, opinions and attitudes is represented in language. (Wei et al., 2013) also supposes that the goal of sentiment analysis is to figure out consumers' sentiments about entities using automatic analysis of text found in reviews. (Dave et al., 2003) used the term 'opinion mining' firstly in an article that was published in the proceedings of the 2003 WWW [1] conference and they discuss that an ideal opinion-mining system would 'analyse a selection of search results for a random item, generating a list of product attributes (quality, features, etc.) and extracting sentiments about each of them (poor, mixed, good)'. (Pang and Lee, 2008) declare that 'the history of the expression sentiment analysis coincides with opinion mining, as many researchers interchangeably used the word 'sentiment' and 'opinion' in their work in regard to the automatic process of analysing texts (Turney, 2002, Sanjiv and Mike, 2001, Tetsuya and Jeonghee, 2003, Anindya et al., 2007). In conclusion, it appears "sentiment analysis" and "opinion mining" represent the same research area. Moreover, there are also other many names and somewhat different tasks, for example, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis and review mining (Liu, 2012).

In general, the textual information is divided into two categories. The first category is the factual information that only holds facts or objective statements about entities. The second category is the subjective information that manifests the actual feeling, opinion of the writer toward entities. In this thesis, we consider the notion of sentiment as the subjective information that shows the feelings or opinions of a person about a specific topic or subject (Turney, 2002).

Sentiment analysis (SA) is an evolving area that triggers the interest of researchers and specifically organisations because SA can be used for decision making. Individuals are no longer restrained to ask opinions from friends about products or services, they can freely come across such information on the Internet. Additionally, organisations may save time and money by avoiding carrying out surveys instead, they can focus on analysing opinions that can be accessed from the Web freely. However, it is important to state that sources that provide opinionated data are noisy, so it is important to only collect the necessary meaning from that information to use it later on.

[1] WWW: World Wide Web

2.5 Depth of sentiment analysis

Based on the levels of granularity (Figure 2.3) of the previous research, sentiment analysis has been mainly studied at four different levels: document level, sentence level, aspect level and user level (Pang and Lee, 2008, Liu, 2012).

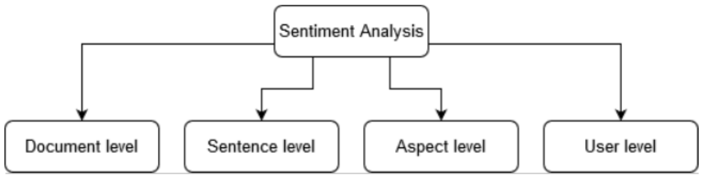


FIGURE 2.3: Different levels of sentiment analysis.

2.5.1 Document level

The task of document-level sentiment analysis is to determine whether a given document conveys an overall positive or negative sentiment. For example, a sentiment analysis system classifies the general polarity of a customer review about a certain product. This level of sentiment classification supposes that one document expresses opinions on a single object, such as customer reviews of products and services, because typically the result of sentiment analysis only has two (positive and negative) or three outputs (positive, negative and neutral). This supposition may be applicable for customer reviews of products and services. But, it may not be suitable for a forum or blog post because in these the author may express sentiments on different products or compare them using comparative sentences.

There are many researchers who have conducted document-level sentiment analysis (Pang and Lee, 2008, Turney, 2002, Balage and Pardo, 2013). They primarily focus on how to isolate the positive texts from the negative texts automatically and they also have proposed different approaches to enhance the accuracy. Due to the simple result of the sentiment classification, the crucial limitation of document-level sentiment analysis is the lack of in-depth analysis (Liu, 2012).

2.5.2 Sentence level

The sentence level of sentiment analysis implicates determining whether each given sentence is expressing a neutral, positive, or negative sentiment (Gamon et al., 2005). There is no key difference between document-level and sentence-level sentiment analysis, because in a way sentences represent short documents (Liu, 2012). It commonly involves two sub-tasks: 1) establishing whether the sentence is a subjective sentence or an objective sentence; 2) if the sentence is subjective, then establishing whether it conveys a positive or negative sentiment (Liu, 2012). This level is associated with the subjectivity classification (Wiebe et al., 1999), which is to filter out the subjective sentences that indicate sentiments or opinions from objective sentences that indicate facts. Thus, it is essential to pinpoint the objective sentences and identify the strength of sentiments. The value of neutral frequently designate the objective sentences. Additionally, it is crucial to state that subjectivity is not equivalent to sentiment, as (Liu, 2012) points out, because objective sentences can also hint sentiments, for example: "I bought this phone one week ago and now the battery only lasts three hours". Those objective sentences that also suggest sentiment also belong to one sub-group of opinionated sentences. This kind of opinionated sentences imply sentiments towards different aspects of object, thus sentence-level classification is not suitable for them (Narayanan et al., 2009).

2.5.3 Aspect level

Classifying opinions at document-level or sentence-level is useful in many cases, but they are inadequate to provide details needed for applications, because they do not spot sentiment targets (Liu, 2012). The aspect level of sentiment analysis focuses on sentiment itself rather than looking at the composition of documents, such as paragraphs, sentences and phrases. It can be decomposed into two sub-tasks: aspect extraction and aspect sentiment classification (Liu, 2012). The task of aspect extraction can also be considered as an information extraction task, which aims to determine the aspects. For example, in the sentence, "My Xiomi screen is beautiful but its battery life is low". We have "Screen" and "battery life" as the aspects of the object "Xiomi". The basic method for extracting aspects is identifying frequent nouns or noun phrases. Then the text holding aspects is classified as positive, negative or neutral (Long et al., 2010). Most of the studies in aspect-level sentiment analysis are based on the assumptions of the pre-defined aspects by keywords (Wang et al., 2011, Li et al., 2015). (Ding et al., 2008) proposed a lexicon-based method for aspect analysis but they suggest that aspects are known beforehand. (Liu, 2012) states that the accuracy of aspect level sentiment analysis is still low because the existing techniques still cannot handle complex sentences well. In conclusion, the aspect level sentiment analysis is harder than both the document-level and sentence-level classifications.

2.5.4 User level

In addition to the previous levels of analysis, some studies conducted sentiment analysis at user-level which consists on studying users' networks and predict users' sentiment based on the sentiment of neighbouring users (Haddi, 2015). Moreover, a number of studies use different models that combine two or three different levels, where the knowledge about one level, for example documents, assists to predict the sentiment of a different level, like sentences (Haddi, 2015).

2.6 Development of sentiment analysis

2.6.1 Text interpretation

Although the research of sentiment analysis and opinion mining have flourished with the rise of Internet and Web 2.0, the previous studies before that also established the foundation for current research. At the early stage of opinion mining extraction, the studies concentrated on text interpretation in some areas such as psychology and politics. Text interpretation can be defined as analysing different formats of texts by using simple computational methods with human assistance to understand the subjectivity, point of view in a given piece of text (Pang and Lee, 2008, Anbananthen and Elyasir, 2013). The research of literary theorist (Banfield, 1982) has been involved in proposing the use of subjective and objective sentences as indicators. According to Benfield's theory, the sentences of narration have been divided into subjective and objective sentences. Subjective sentences represent a character's thought or a perception including the character's emotions, judgments, beliefs, attitudes and affects. Objective sentences only represent the facts, Banfield's theory has been considerably used in the early history of sentiment analysis (Anbananthen and Elyasir, 2013).

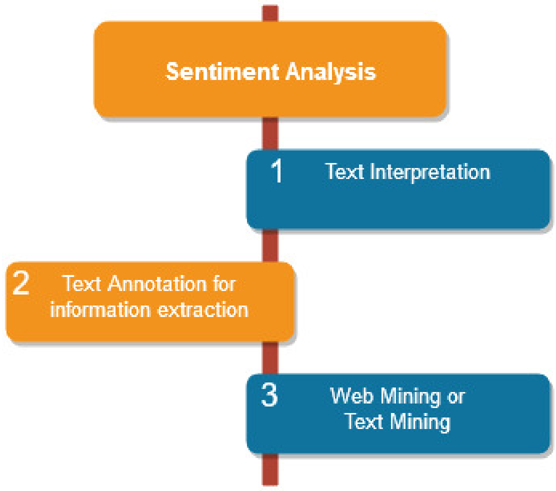


FIGURE 2.4: Evolution of sentiment analysis over the years.

In the light of the growing availability of computer-accessible documents, (Hearst, 1992) implemented one of the first Text-Based intelligent Systems, which provides a search service to answer a broad class of queries for users. Although, it was just a simple 'question-answering system', (Hearst, 1992) deduced that preferably a Text-Based Intelligent System needs to perform full interpretation of any kind of input and allow the results to be retrieved.

2.6.2 Text annotation for information extraction

As seen from the work of (Banfield, 1982) to Hearst's study in (Hearst, 1992), although some researchers have implemented different systems of text interpretation, they involve major human assistance, thus there persisted the difficulty of extracting information from unstructured corpus and performing classification. In order to solve this problem, text annotation was proposed, which is the application of tagging and labelling the corpus by adding new information (attributes) to the text (Anbananthen and Elyasir, 2013, Wiebe et al., 2005). Text annotation research has taken place at many institutions, such as Grenoble, and automated text annotation tools such as footnote and endnote have been utilised in Microsoft Word (Shabajee and Reynolds, 2012). Much of the first studies on information extraction mainly focused on the question answering (Cardie, 1997). Since information extraction is one of the tasks that natural language processing is dedicated to; it is no surprise that the methods of information extraction would be suitable for opinion mining and sentiment analysis. The term opinion-oriented information extraction is used to mention information extraction problems specific to sentiment analysis and opinion mining, which in some cases was abridged to opinion extraction (Pang and Lee, 2008, Cardie et al., 2003). Still, the problem with text annotation is that it demands considerable time and effort to annotate and define the attributes manually in order to generate the corresponding corpus from unstructured text. Therefore, advanced unsupervised learning techniques were proposed for opinion-oriented information extraction to work on different levels of opinion classification in order to improve the accuracy of classification (Anbananthen and Elyasir, 2013). One of the important approaches in text annotation for information extraction is part-of-speech tagging (POS), which is where each word present in the text is individually classified as one of the parts of speech such as noun, verb, adjective, preposition, adverb, conjunction, etc., thus it can provide an important amount of information about the word and its neighbours (Jurafsky and Martin, 2008). (Dave et al., 2003) have implemented a method for automatically differentiating between positive and negative reviews by training a classifier using a corpus of self-labelled reviews from websites. (Yi et al., 2003) proposed a sentiment analyzer (SA) that extracts sentiments about a subject from online text by detecting all the online references to the given subject. The sentiment analyzer is one of the first endeavours on sentiment analysis which made use of the bag of words (BoW) (Anbananthen and Elyasir, 2013).

2.6.3 Web mining or Text mining

Due to the evolution of the Internet and the rise of machine learning methods in natural language processing, the real sense of sentiment analysis began when it had been involved in web applications, apart from other domains such as politics (Anbananthen and Elyasir, 2013). Sentiment analysis has been used to investigate and analyse text such as blog posts, social networks comments and online product reviews. Several new techniques and applications were introduced for sentiment analysis in the era of Web 2.0. (Liu et al., 2005) implemented an analysis system called Opinion observer to classify positive and negative product reviews from online reviews and compare consumer opinions. The results of the proposed system represent the strengths and weaknesses of each product.

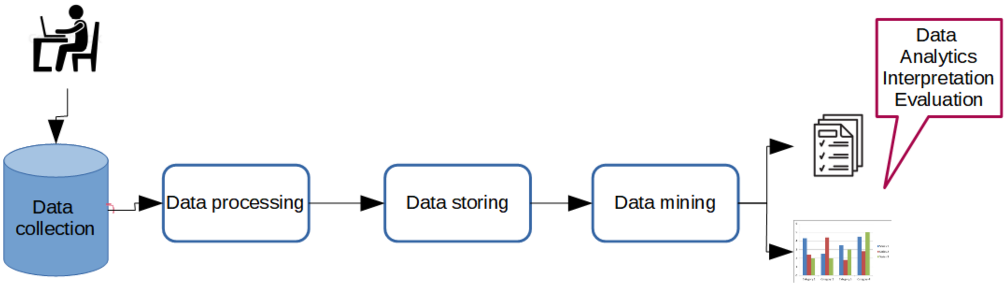


FIGURE 2.5: The key steps in the Web mining process.

The work of (Kim and Hovy, 2006) focused on another challenging problem in sentiment analysis, which is to identify the reasons behind opinions in online reviews. They proposed a system made for extracting the pros and cons automatically from online reviews. The evolution of opinion mining is not only just about the unceasing development of methodologies, but also its application in various domains. After the era of Web 2.0, sentiment analysis has been applied extensively in many fields such as marketing, e-commerce, health care, politics and education with aim of improving products and services (Esuli and Sebastiani, 2006, Khan et al., 2016, Anbananthen and Elyasir, 2013).

2.7 Approaches of sentiment analysis

Sentiment analysis also referred to as opinion mining has been the center of growing attention in research and business industry due to its massive value and potential for practical applications, particularly in the era of Web 2.0 (Pang and Lee, 2008, Montejo-Ráez et al., 2014). Traditionally, sentiment analysis is considered as a binary classification problem of sentiment (Dave et al., 2003, Pang et al., 2002). (Esuli and Sebastiani, 2005) explained that sentiment classification can be divided into three main sub-tasks, which are illustrated in the Figure 2.6:

- **Task 1:** Determining subjectivity, as in deciding whether a given text contains factual information or subjective information;
- **Task 2:** Determining the orientation or polarity of the text, as in deciding whether a given subjective text holds a positive or negative opinion;

- **Task 3:** Determining the strength of that orientation;

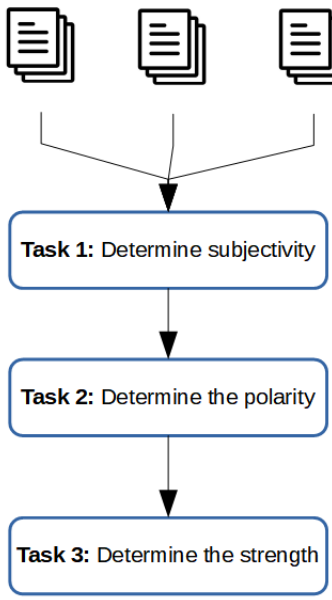


FIGURE 2.6: Sentiment analysis sub-tasks.

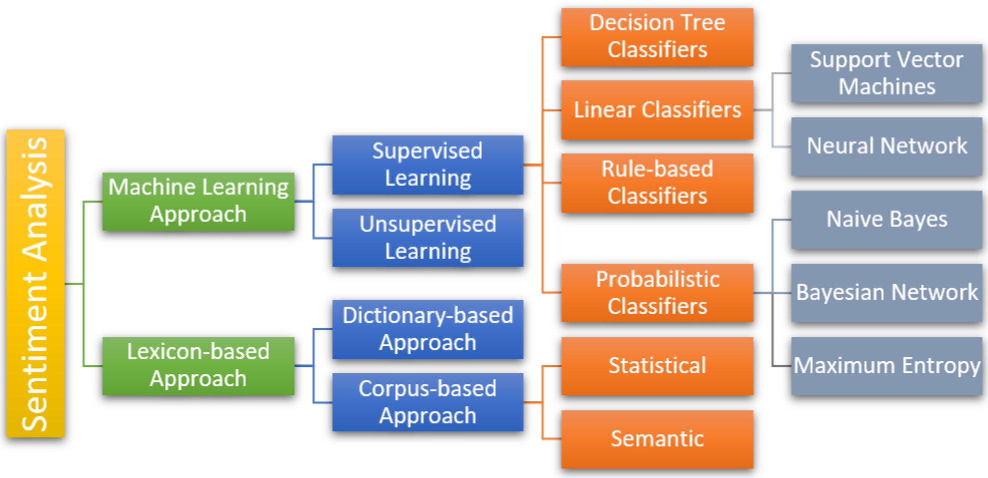


FIGURE 2.7: Sentiment analysis approaches.

Due to the enormous volume of subjective data found on the Web, automated sentiment analysis is required to tackle this problem (Liu, 2010). As shown in Figure 2.7 above, research on sentiment analysis has been dominated by two main approaches: semantic orientation approach and machine learning approach (Chaovalit and Zhou, 2005, Medhat et al., 2014). The semantic orientation approach is also known as lexicon-based approach in some studies and it considers words and phrases as indicators of semantic orientation and the overall polarity of the text is generated by calculating the averaged sum of all polarities (Pang and Lee, 2008, Hatzivassiloglou and McKeown, 1997). The other approach machine learning is based on the implementation of machine learning algorithms. Its main focus is determining the suitable machine learning algorithm and using the right text features to classify the polarities of the text (Pang et al., 2002, Brooke, 2009). Other researchers also refer to these two approaches as supervised learning and unsupervised learning (Liu, 2012, Khan et al., 2016, Montejo-Ráez et al., 2014). The current work seek to exploit the advantages of both techniques. Moveover, by combing both methods, hybrid classification systems of sentiment analysis were also implemented (Prabowo and Thelwall, 2009).

2.7.1 Semantic orientation approach

The semantic orientation or polarity of a word is the characteristic that designates the direction of the opinion. Semantic orientation has various directions (positive, negative or neutral) and different intensities from mild to strong (Turney and Littman, 2003). Positive semantic orientation of a word indicates a desirable state (e.g., beautiful, good), while negative semantic orientation of a word conveys an undesirable state (e.g. hate, ugly) (Hatzivassiloglou and Wiebe, 2000). The studies proclaim that words with polarities, in particular adjectives, are considered as good indicators of subjectivity (Turney, 2002, Hatzivassiloglou and McKeown, 1997). Therefore, the semantic orientation approach focuses on words and phrases as the holders of polarities, and the overall semantic orientation of the entire text is calculated by the sum of indicators with polarities. It is also called the lexicon-based approach in some works (Liu, 2010, Medhat et al., 2014, Brooke, 2009). The prototype of semantic orientation approach for sentiment analysis is the lexicon-based method, which uses a dictionary of sentiment words with their assigned polarities and intensity or strength to identify the sentiments in the corpus to be analysed (Taboada et al., 2011). These sentiment words are referred to as opinion words, which are frequently used to indicate positive or negative sentiments. For example, words such as beautiful, amazing, good are positive sentiment words while on the other hand words such as disgusting, horrible, bad are negative sentiment words. Commonly, the dictionary of sentiment words is also known as sentiment lexicon or opinion lexicon (Liu, 2012).

There are three different techniques to produce the dictionaries for the lexicon-based approach: 1) the dictionary of sentiment words can be generated manually, which is more precise but time consuming and expensive (Brooke, 2009, Stone et al., 1966, Hu and Liu, 2004); 2) the dictionary can be produced depending on the co-occurrence patterns in a large corpus, which is known as corpus-based approach (Turney, 2002, Hatzivassiloglou and McKeown, 1997, Turney and Littman, 2003); 3) another method is based on the bootstrapping using a selection of seed opinion words and an online dictionary such as WordNet for sentiment analysis, which is called dictionary-based method (Pang and Lee, 2008, Liu, 2012, Kim and Hovy, 2004).

2.7.2 Corpus-based method for sentiment analysis

Seeking to create a sentiment dictionary, Bing Liu explained that corpus-based approach can be performed in two situations. First one is a recognition of opinion words and their assigned polarities in the domain corpus using a given set of opinion words. The second one is for creating a new lexicon within the specific domain from another existing lexicon. The results obtained so far indicate that even if opinion words are domain-specific there is a possibility that the same word will have totally different orientation depending on context. The corpus-based technique is primarily dependent on the syntactic or co-occurrence patterns and a list of seed words to enlarge the selection of opinion words (Liu, 2012, Luo et al., 2013). Seed words are a small set of words with powerful positive or negative orientation, which are frequently defined and collected manually. The key idea was first introduced by (Hatzivassiloglou and McKeown, 1997), where they presented an approach for automatically predicting the semantic orientation of words. They demonstrated that conjunctions between adjectives provide additional information about their polarity. For example, the two adjectives linked by "and" usually are of the same orientation: clever and beautiful; those linked by the conjunction, such as "but", usually denote opposing orientations: smart but arrogant. Their approach reported satisfactory results. Following this idea came the work of (Turney, 2002) which is a famous example of sentiment analysis via his corpus-based approach and technique of combining the use of mutual information and co-occurrence in the text has been employed by a number of researchers (Pang and Lee, 2008). His method is described in the following stages:

Firstly, sentences containing adjectives or adverbs are used because adjectives are good indicators of subjectivity (Hatzivassiloglou and Wiebe, 2000, Wiebe, 2000). However, (Turney, 2002) states that an individual adjective or adverb may convey subjective feelings or emotions but it may not be able to give sufficient context to identify its semantic orientation. For example, the adjective 'soft' may have a positive orientation in a review about beds but it could totally have opposite polarity in a phone case review. Thus, two successive words are extracted from the given text: one word needs to be an adjective or an adverb, then POS tagger has been applied to the given text. Secondly, the polarity of extracted sentences would be determined by using the Pointwise Mutual Information (PMI) algorithm. PMI algorithm is used to determine the associations between the phrases and the two seeds word: 'excellent' and 'poor'. These particular words were picked out because in a five star review system, the word 'excellent' is commonly used to express five star review while the word 'poor' is used in one star review. The following formula is used to calculate the semantic orientation of each extracted phrase:

$$SO(phrase) = PMI(phrase, 'excellent') - PMI(phrase, 'poor') \tag{2.1}$$

According to the above formula (Turney, 2002), if the phrase leans to be more strongly linked with the word 'excellent' than 'poor', the SO(phrase) would be positive; if not the semantic orientation of the phrase would be negative, if it is more strongly linked with the word 'poor'. Thirdly, after determining the SO values of all extracted phrases in the given text, the polarity of the text is the average of SO values of all phrases. The overall text is classified as positive, if the average SO is positive, if not negative. Based on this work (Turney and Littman, 2003) proposed a somewhat modified approach to calculate the semantic orientation of words by using the same PMI algorithm. Rather than only using two sentiment words ('excellent' and 'poor'), they expanded the set of words to fourteen (seven words in each polarity). They reported that their method has achieved 82.8% in terms of accuracy. By following the insights of the co-occurrence approach proposed by (Hatzivassiloglou and McKeown, 1997), (Rice and Zorn, 2013) built a corpus-based dictionary. They tested the dictionary using the Cornell Movie Review Data introduced by (Pang and Lee, 2004) and report the accuracy of their measure is 72.5%, which comes near to matching the accuracy of machine learning classifiers.

The corpus-based method has a key advantage in that it is capable to deal with domain and context specific opinion words and their polarities (lexicons), but it requires a large corpus to cover all words, which is difficult to develop. Altogether, the performance of lexicon-based approaches regarding time complexity and accuracy primarily depend on the number of words present in the dictionary, that is, performance decreases significantly with the exponential growth of the dictionary size (Malik and Kumar, 2018). In conclusion, it is not as powerful as dictionary-based techniques for sentiment analysis due to the restriction of words that compose the corpus (Liu, 2012, Luo et al., 2013).

2.7.3 Dictionary-based method for sentiment analysis

The dictionary-based method has taken a slightly different path for sentiment analysis, which depends on a dictionary (lexicon) to collect opinion words, which is also mentioned as lexicon-based method in some previous works (Pang and Lee, 2008, Montejo-Ráez et al., 2014). (Liu, 2008) states that it a straightforward and efficient approach to produce the sentiment lexicon. The dictionary-based approach begins with assembling a small selection of words with different polarities manually, followed by utilising an existing dictionary (e.g. WordNet) to expand the set of opinion words via their synonyms and antonyms. The newly extracted words from the dictionary are added to the first selection of opinion words. In the end, manual investigation is needed for possible correction before using the constructed sentiment lexicon for sentiment analysis (Hu and Liu, 2004). There are a number of studies that have taken on this technique along with an online dictionary for sentiment analysis (Kim and Hovy, 2004, Montejo-Ráez et al., 2014, Hu and Liu, 2004, Kamps et al., 2004, Mohammad et al., 2009, Poria, 2012). (Hu and Liu, 2004) have focused their work on the classification of customer reviews, specifically they collected product features that hold sentiments, then classified sentences based on those features. In order to give a positive or negative label to a sentence, first, they extracted adjectives from each review. The classification was based on the polarity of an adjective, their method showed good results with an accuracy of 84%.

Opposed to the technique that only can identify the polarity of adjectives (Kim and Hovy, 2004, Hatzivassiloglou and McKeown, 1997, Hu and Liu, 2004). The work of (Esuli and Sebastiani, 2005) has taken a new step in sentiment analysis with the birth of SentiWordNet (Esuli and Sebastiani, 2006, Baccianella et al., 2010). It is an extension of WordNet where each term is labelled with sentiment orientation information. SentiWordNet is constructed in a two-step stage: firstly, an initial set of words both positive and negative is collected manually and then WordNet is used to extract synonyms and antonym to generate more words. Secondly, the selection of terms is used to train the classifier. The Random Walk classifier is used in sentiment analysis based on the assumption that *'if the classification process starts at a given word it is more probable to come across another word with the same semantic orientation before a word with a different polarity'* (Montejo-Ráez et al., 2014).

The work of (Taboada et al., 2011) investigated various dictionaries to compare their performances. Their results indicated that SentiWordNet gave better performance than the Google dictionaries, Maryland dictionaries and the General Inquirer lexicon (Stone et al., 1966). (Jain and Pandey, 2013) have implemented a method by using the lexical resource SentiWordNet to assign the polarity at sentence level. They used part of speech tagger as feature for sentiment calculation. Their results obtained an average accuracy of 69.1%. By making use of SentiWordNet, (Khan et al., 2016) presented a hybrid method to predict the sentiment orientations of movie reviews. They constructed a sentiment dictionary SentMi and implemented a supervised learning algorithm for sentiment classification. The results of their proposed hybrid method and its comparison with SentiWordNet classifier are reported: their method achieved an average accuracy of 76%, while SentiWordNet obtained an accuracy score of 68%.

Previous research illustrated that dictionary-based methods performed very well in different domains (e.g. movie reviews, tweets) against traditional methods. However, dictionary-based technique has several disadvantages. The main issue resides on the fact that the process of collecting synonyms and antonyms is time-consuming. Additionally, these dictionaries contain only formal words, but social media text such as tweets or reviews are full of informal language. Overall, the main drawback of dictionary-based approach is the incapacity to identify sentiment words with domain and context specific polarity orientations (Malik et al., 2018).

2.7.4 Linguistic approach for sentiment analysis

The linguistic approach is based on the syntactic properties of words, phrases, negations, and the structure of the text to predict the text orientation. This method is commonly combined with a lexicon based method (Turney, 2002, Tan et al., 2011, Thelwall et al., 2011). One of the techniques used in the linguistic approach is based on Parts-Of-Speech (POS). This feature defines the syntactic patterns or categories of the words. To determine these patterns, n-grams are used where; an n-gram is a sequence of n words from a given sequence. Usually, we can use uni-grams, bi-grams, tri-grams, and n-gram for more than three words.

In conclusion, the three different approaches we surveyed can be applied individually or combined together. For example, machine learning and linguistic approaches can be combined (Montejo-Ráez et al., 2014), so that the selected features for training are POS tags. The lexical based approach can be combined with a linguistic approach, so the lexicon is produced for example from the adjectives that appear in a given text or in a specific domain. Afterwards, those adjectives can be labelled in a lexicon as positive or negative. To illustrate, the words 'beautiful' and 'ugly'. In any case does this suggest to ignore the importance of other parts of speech like verbs or nouns, given some of them express very strong sentiment, such as the verb 'hate' (Pang and Lee, 2008).

2.8 Machine learning approach

The advancement of machine learning algorithms in natural language processing (NLP) has led to increased popularity of research in sentiment analysis. In the approach of machine learning, a feature representation of text is used combined with several classifiers such as naive Bayes, Support Vector Machines (SVM), Maximum Entropy, which are frequently utilised to build the classifiers for sentiment analysis. These classifiers can learn the features or decision basis of sentiment classification based on training data, then they are used to apply sentiment analysis automatically (Waila et al., 2012, Ghiassi et al., 2013). This obviously shows that the machine learning approach for sentiment analysis is fairly a supervised learning framework, where a large number of labelled training data are needed to train the classifier before it is used for classifying the new data later on (Pang et al., 2002, Waila et al., 2012). A detailed discussion on how these machine learning algorithms work are beyond the scope of this thesis. Nevertheless, the idea behind the machine learning approach for sentiment analysis is straightforward. It is based on the framework of supervised classification (Figure 2.8) and composed of two stages:

1. Learning the model from a corpus of labeled training data using different features;

2. Classifying the new data based on the trained model. Overall, the whole classification process involves several sub-tasks: data pre-processing, feature selection, data representation, data classification and evaluation (Khairnar and Kinikar, 2013).

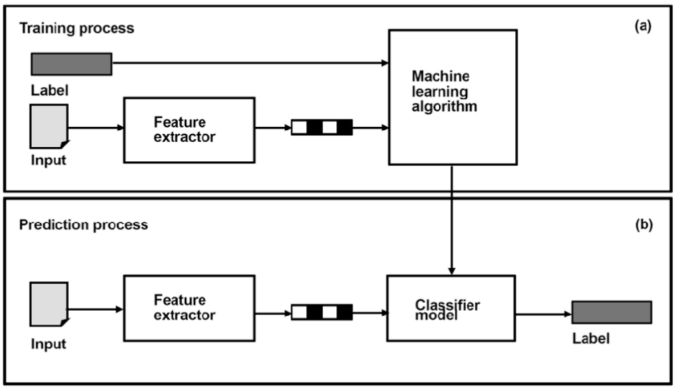


FIGURE 2.8: Standard framework for supervised classification.

2.8.1 Sentiment analysis using supervised learning

Sentiment classification can clearly be represented as a supervised learning problem with two class labels (positive and negative). Training and testing data used in research are for the most part product reviews, which is not surprising due to the above mentioned assumption. Since each review on a regular review site already has a reviewer attributed rating (e.g., 1-5 stars), training and testing data are easily and publicly available. Typically, a review with 4-5 stars is identified a positive review (thumbs-up), and a review with 1-2 stars is considered a negative review (thumbs-down) (Liu, 2010). In sentiment classification, sentiment or opinion words that convey positive or negative sentiments are important, e.g., great, excellent, amazing, horrible, bad, worst, etc. The application of existing supervised learning approaches to sentiment analysis can be pretty straightforward e.g., naive Bayes, and support vector machines (SVM), decision trees (DT), logistic regression (LR), etc. Research were carried out where various kinds of features and techniques were used for learning and training classifiers (subsection 2.8.3). In most machine learning applications, the main task of sentiment classification is to design a suitable set of features.

Some of the features used in research are described in detail (subsection 2.8.3). For a more thorough survey of existing features, please refer to (Pang and Lee, 2008). Aside from classification or prediction of positive or negative sentiments, studies have also been conducted on predicting the rating scores (1-5 stars) of reviews (Pang and Lee, 2005). Another compelling research direction that has been investigated, is the transfer learning or domain adaptation as it has been indicated that sentiment classification is extremely sensitive to the domain from which the training data are extracted. A classifier trained using opinionated texts from one domain commonly performs poorly when it is experimented with or tested on opinionated texts from a different domain. This issue is caused by the fact that words used in different domains to express sentiments can be considerably different. To make matters worse, the same word may designate positive sentiment in one domain, but in another domain may indicate a negative one. Here the classic example is from the work of (Turney, 2002), where the adjective unpredictable can mean a negative sentiment in a car review (e.g., "unpredictable steering"), but it could have a positive sentiment in a movie review (e.g., "unpredictable plot"). Therefore, in cases similar to the mentioned example domain adaptation is necessary. Many research has been proposed to deal with domain adaptation, where they used labelled data from one domain and unllabeled data from the target domain and a selection of opinion words as features (Blitzer et al., 2007, Aue and Gamon, 2005, Yang et al., 2006).

2.8.2 Sentiment analysis using unsupervised learning

In unsupervised learning, unllabeled datasets are used to uncover the structure and identify the similar patterns from the input data. This method is usually used when there is difficulty in collecting reliable labelled dataset, yet collecting unllabeled data is easier. It does not cause any difficulties when new domain-dependent data have to be retrieved. It is obvious to imagine that opinion words and phrases are the essential indicators in the sentiment classification process. Therefore, the application of unsupervised learning based on such words and phrases would be quite common. (Turney, 2002) proposed an unsupervised machine learning approach for the classification of reviews. Typically, reviews are classified into recommended (thumbs up) and not recommended (thumbs down). The author employed POS to the document in order to decide which phrases have to be extracted. In this work, phrases containing adjectives, adverbs, verbs or nouns were retrieved. Afterwards, the semantic orientation of each phrase from the review was calculated. Then, Pointwise Mutual

Information (PMI) was applied to find semantic orientation. PMI measures the semantic resemblance of a given phrase with the following two words: "excellent" and "poor", where if the phrase has a bigger association with "excellent" then it is positive and if its association with "poor" is bigger then negative. Last step is to identify the sentiment for the whole review and conclude whether it is recommended or not recommended. If the average semantic orientation score is positive then the review is determined as recommended and not recommended otherwise. The average accuracy reported was 74%.

2.8.3 Feature Selection

Engineering the best feature set for sentiment analysis has the highest importance as it has a massive influence on the evaluation results. This section presents the most frequent features and these are commonly pre-processed beforehand using different techniques in order to reduce the feature space. Assembling the feature set is necessary when building a model in order to reduce overfitting, improve accuracy and reduce training time. In general, feature selection is considered as the important part of dealing with the corpus' training data in the machine learning approach (Kummer and Savoy, 2012, Basant and Namita, 2016). This part consists of converting a piece of text into a feature vector or other existing representations for computational processing (Pang and Lee, 2008).

As a start, the training data are labelled as positive, negative or neutral and next a selection of features is extracted from the labelled training data. Afterwards, the extracted features can be represented using simple value types, such as Booleans, numbers and strings. For example, the presence or absence of words that appear in the text can be considered as features. Since the training data is normally composed of two groups (positive and negative), each individual word in each group can be considered as a feature vector. Some words such as stopwords (e.g. "a", "is", "the") and some alphabet do not convey any information about sentiment therefore, these words are usually filtered out or removed. Adding single words to the set of features is a technique frequently known as uni-grams method (Pang et al., 2002). There is considerable research dealing with the feature selection in machine learning approach, however an in-depth discussion of such work is beyond the extent of this thesis. Nevertheless, some typi cal features used in sentiment analysis are listed below as well as a brief description of each one (see Table 2.1).

TABLE 2.1: Examples of different features used in machine learning.

Feature	**Description**
Terms and their frequency	Words and their frequency of appearance in text. The TF-IDF method is applied
Part of Speech	It is used because adjectives are considered as important indicators of sentiment.
n-grams	The position of a word can possibly influence the sentiment.
Syntactic dependency	It is usually generated with existing lexicons.
Negation	Negation words are important because they alter the overall sentiment.
Opinion words	Words used to express positive/negative sentiments.
Punctuation	Special characters like ! ? ' " or/.
Emoticons	Lists of positive, negative emoticons in the text.
Orthographic Features	It is based on the appearance of the word,e.g. the first letter is a capital letter.
Word2vec	Linguistic resource for word embeddings, with large input and produces vectors.
Glove	It is an unsupervised learning algorithm for obtaining vector representations for words.

The key purpose of feature selection is to decrease the dimensionality of the feature space and hence, make the computational processing easier. A number of studies (Pang et al., 2002, Kummer and Savoy, 2012, Abbasi et al., 2008) have tried to experiment with different feature types using various algorithms in order to enhance the performance of sentiment classification. Various combinations of features extracted in the machine learning approach could generate to different performances. For example, the problem whether n-gram feature can improve the accuracy of sentiment classification has been heavily investigated by different researchers. (Pang et al., 2002) have evaluated a number of features, such as uni-grams and bi-grams, with and without POS labels. The results showed that the SVM classifier achieved better sentiment classification using uni-grams as feature rather than bi-grams. In contrast, (Dave et al., 2003) carried on an experiment to classify product reviews collected from the websites CNET[2] and Amazon[3] into two polarities (positive and negative) and surprisingly, they concluded that bi-grams and tri-grams gave better results than uni-grams in some cases.

[2]https://www.cnet.com/
[3]https://www.amazon.com/

(Ng et al., 2006) conducted a comparison study of the n-gram features by classifying the polarity of a review as positive or negative. Primarily, they trained a first polarity classifier using only uni-gram and following that, they trained a second polarity classifier based on all uni-grams, bi-grams and tri-grams. The results reported the accuracy of the first classifier was 87.2% and 79.2% for the second classifier. N-gram features provide an easy way to capture the context, if n is too small, the model does not have enough context. Oppositely, if n is too large *(for $n > 3$)*, it is very hard to computationally generate the features and also it will produce serious sparse data issues (Peng et al., 2003, Ng et al., 2006, Bespalov et al., 2011).

2.8.4 Evaluation metrics

After the model for sentiment analysis is trained and tested, its performance evaluation is mandatory. The performance of the classifiers used for sentiment analysis is measured by calculating various metrics such as accuracy, precision, recall and F-measure also referred to as F-score or F1 score. Below, we will describe these measures on a binary classification of positive and negative labels, but normally any number of labels can be used. We can show the results in the form of a confusion matrix.

- Positive (P) - positive text classified as positive.
- Negative (N) - negative text classified as negative.
- False positive (FP)- negative text classified as positive.
- False negative (FN) - positive text classified as negative.

TABLE 2.2: Confusion matrix.

	Positive	**Negative**
classified as positive	Positive (P)	False positive (FP)
classified as negative	False negative (FN)	Negative (N)

Now we can straightforwardly define the other metrics; accuracy, precision, recall and F-measure as follows:

$$Accuracy = P + N/P + N + FP + FN \tag{2.2}$$

$$Precision = P/P + FP \tag{2.3}$$

$$Recall = P/P + FN \tag{2.4}$$

$$F - measure = 2P/2P + FP + FN \tag{2.5}$$

Accuracy is a percentage of all correctly predicted labels in contrast to all sentences. Precision is a measure of trust, that the items identified as positive are really positive. Recall is a measure of trust, that all the positive entities are identified. F-measure is a harmonic mean between precision and recall and it is considered to be an overall proportion.

2.9 Supervised learning methods for SA

After a collection of features is determined from the training data, an algorithm is implemented to learn those features. If a specific feature leans to be always true when the training data belongs to a particular class (positive, negative), the used classifier will learn that this feature is a good indicator of that class. For example, if the word "good" appears consistently in positive training data, the algorithm will learn the feature word "good" is a useful indicator of positive orientation. When the classifier is used to test the new data (testing data), it will extract values based on the features from the new data and multiply those values by the weights previously learned from the training data. The sum of the value will display the results of the sentiment classification (Pang and Lee, 2008, Liu, 2010).

These last years, a lot of work has been done in the field of data analytics precisely sentiment analysis in natural language and social media posts. To determine whether a piece of text expresses a positive or negative sentiment, different approaches are commonly used such as support vector machines (SVM), logistic regression (LR), maximum entropy (ME)...etc. However, most popular classifiers in the field of sentiment analysis are: Naive Bayes (NB), Reccurent Neural Networks (RNNs), SVM and Convolutional Neural Networks (CNN) also known as ConvNets are a deep learning tool that has gained expertise in computer vision applications (Srinivas et al., 2016). It was first introduced in (LeCun et al., 1989) to recognize handwritten ZIP code in 1989, (LeCun et al., 1998) later extended to recognition and classification of various objects such as hand-written digits (MNIST), house numbers (Semanet et al., 2012), Caltech-101[4] (Fei-Fei et al., 2007), traffic signs (Semanet and LeCun, 2011), and during recent years the work of (Krizhevsky et al., 2012) 1000-category ImageNet[5] dataset.

(Go et al., 2009) implemented the same algorithms used in the work (Pang et al., 2002) to classify the sentiment of Twitter messages, in which the training data were a selection of tweets with emoticons. Their experimental results reported that the naive Bayes, SVM and Maximum Entropy classifiers respectively obtained the accuracy of 81.3%, 82.2% and 80.5% corresponding to the reported results in the study of (Pang et al., 2002). By following the work of (Go et al., 2009), (Pak and Paroubek, 2010) had taken on a study on the credibility of Twitter as a source of data used for sentiment analysis. They described a method for constructing a corpus of tweets that were later on used for training the classifier and then to determine the polarity of the tweets. They implemented three different algorithms: naive Bayes, SVM and Conditional Random Fields (CRF). Their results revealed that the best configuration for the sentiment analysis on Twitter was using naive Bayes classifier along with n-grams and POS tags as features.

A framework to detect the polarity product reviews was explored by (Lin et al., 2012), in which the scores of sentiment words were assigned. They conducted their experiment using SVM in order to predict the polarity of product reviews from four different domains: books, electronics, DVDs and kitchen appliances. The reported accuracies varied from 73% to 88% depending on the selected features and domain application. The results revealed that the performance of the classifier gave higher accuracy in the domains of electronics and kitchen appliances than those in books and DVDs domains. The best reported performance was using uni-grams as feature and in the domain of kitchen appliances with 88% accuracy score.

The choice of using neural networks to build natural language processing (NLP) applications has attracted growing interest in research community and they are regularly applied to all NLP tasks (Kim, 2014). The principal idea of convolutional neural networks is to think about feature extraction and classification as one combined task. (Dhande and Patnaik, 2014) proposed a sentiment analysis approach combing neural network (NN) algorithm with naive Bayes for NN classifier is able to handle the correlation and dependence. In their experiment, they implemented three different

[4]http://www.vision.caltech.edu/Image-Datasets/Caltech101/
[5]http://www.image-net.org/

classifiers with uni-grams as feature to analyse 2000 movie reviews and their results are presented in Table 2.3.

In the work of (Ouyang et al., 2015) a CNN and word2vec methodology was proposed for movie reviews sentiment analysis using a dataset from rottentomatoes.com[6]. The dataset consisted of 11,855 reviews with five different sentiment classes: negative, somewhat negative, neutral, positive and somewhat positive. Their CNN model used three different convolution layers with different kernels and each layer was followed by a dropout layer, and normalization layers. For result evaluation, they compared their model against other models including naive Bayes, SVM, Recursive Neural Network (RNN) and Matrix-vector RNN (MV-RNN). The result shows that performance is best when it comes to classifying every review into the five different classifications. Their model achieved a test accuracy of 45.4% on the test dataset.

TABLE 2.3: Accuracy of classifiers in the above mentioned work.

Authors/year	**Method/feature**	**Best results**
Go et al. (2009)	NB, SVM and ME with n-grams	80.4%—82.9%
Lin et al. (2014)	SVM with n-grams	73%—88%
Dhande and Patnaik (2014)	NN-NB with n-grams	80.65%
Severyn et al. (2015)	CNN with weights	84.97%
Ouayang et al. (2015)	CNN with word2vec	45.4%
Houshmand (2017)	NN, NB with word2vec	40.5%—46.4%

In the work of (Houshmand, 2017) a comparison between different neural networks architectures against the naive Bayes algorithm to see how well they performed on movie reviews from the Stanford Sentiment Tree bank dataset. The results of their study showed similar accuracy between the neural networks used (recurrent, recursive and convolutional neural networks) and naive Bayes. One interesting thing about the result was the fact that their model's accuracy improved significantly by adding a word vector from word2vec to the network. Their model reached an accuracy of 46.4% on the test data while the CNN without a word vector got 40.5%

Recently, the biggest reason to adopt convolutional neural networks in natural language processing, sentiment analysis, text, topic and document classification is due to the following key reasons; 1. CNN can extract an area of features from global information; 2. It is able to consider the relationship among these features (Kim, 2014); 3.

[6]https://www.rottentomatoes.com/

Text data features are extracted piece by piece and the relationship among these features, with the consideration of the whole sentence, thus, the sentiment can be understood correctly. The scope of using this methodology in data analytics has proven to be advantageous in various ways.

2.10 Domains of sentiment analysis

2.10.1 Twitter Sentiment Analysis

Research on acquiring and analysing Twitter data has been growing exponentially over the past decade. This interest is due to several reasons: First, social media text is a rich source of public information. Second, information on Twitter is open, has clear character number and well-documented. Third, it is in real-time where the appearance of messages occurs, figuratively speaking, with the speed of thought. Fourth, this site is used to pass and express opinions by different sectors. Twitter content analysis can help to evaluate changes in moods of many users, to reveal their preferences, likes and dislikes.

Posts on Twitter referred to as tweets are considered as a gold mine for sentiment analysis researchers. There is empirical evidence that Twitter users tend to share posts (tweets) or opinions about products or services (Pak and Paroubek, 2010), and these tend to be short (at most 280 characters long), informal and normally straight to the point. Sentiment analysis tasks that can be applied to Twitter data are polarity classification and sentiment identification. Considering the short nature of tweets, a sentence-level classification approach can be conducted, based on the assumption that tweets express sentiments about one single entity. Furthermore, retrieving messages from Twitter is a pretty straightforward process using the public Twitter API[7].

In (Go et al., 2009), the Twitter API was fetched to extract tweets containing both positive and negative emoticons for building a training dataset for their maximum entropy classifier. A good accuracy of 83.7% was obtained when the model was trained on a manually labelled dataset. They experimented with uni-grams and bi-grams as features. Moreover, they tried to reduce the feature sparse by replacing repeated letters (e.g., huuungry to hungry, loooove to love) and replacing users tags found in tweets by the '@' symbol. Just the same for URLs were replaced with the symbol 'URL'.

[7] https://developer.twitter.com/en/docs

Pak and Paraoubek conducted similar work to their previous research in (Pak and Paroubek, 2010). In addition to positive and negative classes extracted from emoticons, they added a neutral class retrieved from tweets posted by accounts of popular newspapers and magazines such as the New York Times. The corpus was used to train the classifier and perform 3 class sentiment analysis. This work aimed on taking advantage of both types of labels, emoticon-based and human-labelled. In (Zhang et al., 2011), the authors proposed a lexicon based method to apply sentiment analysis on Twitter data. They trained a classifier using a selection of annotated tweets. Their method is able to capture domain specific sentiment patterns and achieved high precision, yet low recall.

(Speriosu et al., 2011) proposed another supervised learning approach combined with a lexical approach. The authors constructed a graph consisting of users, tweets, words, hashtags and emoticons as its nodes. A sub-selection of these nodes is labeled beforehand with a score from the polarity lexicon. These labels are distributed throughout the built graph. An exhaustive survey of approaches utilising unlabelled data for Twitter sentiment analysis based on self-training, co-training, topic modelling, and distant supervision is provided in (Silva et al., 2016).

In (Kouloumpis et al., 2011) demonstrated the impact that feature selection has on the model's performance. At first, they trained their model using different linguistic features in addition to existing lexical resources developed for micro-blogging to detect the sentiment orientation. Their results showed that POS features and features from existing sentiment lexicons were effective. Additionally, they also concluded that the training data will be of less gain if they consider micro-blogging features (Nahili et al., 2020).

In 2013, The Semantic Evaluation (SemEval) workshop held the "Sentiment Analysis in Twitter task 3" with the goal of encouraging research in social media sentiment analysis (Rosenthal et al., 2014). The competition was divided into two sub-tasks: the sentence level and the message level. The team that reported the best performance in both tasks was the NRC-Canada team (Mohammad et al., 2013). They implemented a supervised learning method based SVM classifiers using the following selection of features: n-grams, part-of-speech tagging, the number of words with repeated characters, the number of words in uppercase, presence of positive or negative emoticons, the number of continuous sequences of dots, question marks and exclamation marks. Additionally, they used features derived from their previously proposed polarity lexicons (NRC-Hashtag [8] and Sentiment140 [9] lexicons), which were built using a large selection of tweets.

Following the above research was the work of (Goncalves et al., 2013) in which an ensemble approach combining different sentiment analysis techniques for polarity classification of social media messages. The authors evaluated the methods in regards to their corresponding classification performance, proving that their combination achieved a better percentage of correctly classified messages. For their sentiment analysis research, (Severyn and Moschitti, 2015) developed a convolutional neural network for analysing tweets. The representation of each tweet was using a matrix whose columns correspond to the words in the tweet, by keeping track of the order in which they appear. The words are represented by dense vectors or embeddings trained from a large corpus of unlabelled tweets. The network is build consisting of the following layers: an input layer with the specified tweet matrix, one convolutional layer with a rectified linear activation function, followed by a max pooling layer, and a softmax classification layer. The parameters of the neural network are pre-trained using emoticon-annotated data, and then trained with the hand-annotated tweets from the SemEval competition. Experimental results proved that the pre-training phase leads to suitable initialisation of the network's parameters, and therefore, had positive affect on the classification accuracy.

(Debjyoti et al., 2017) investigated a data science project called US Election 2016[10], where they performed a spatial-temporal sentiment analysis from large-scale social media data. Their goal was to discover sentiment on Twitter towards either the democratic or the republican party at US county and state levels over any arbitrary temporal intervals. They used a huge set of geo-tagged tweets from a period of 6 months before the US Presidential Election in 2016. Their results showed effective outcomes and proved that by integrating and developing a combination of machine learning and data management techniques, it is possible to perform a spatial temporal sentiment analysis at scale. The adaptation of their results towards solving and influencing other interesting social issues such as building neighborhood happiness and health indicators prove its potential.

[8] https://saifmohammad.com/WebPages/lexicons.html
[9] http://help.sentiment140.com/
[10] https://dl.acm.org/doi/10.1145/3097983.3098053

In the work of (Mohamad et al., 2017), sentiment analysis was applied to analyse and extract the polarity of sentiment from product reviews (laptop and restaurant) collected in the SemEval 2014 Task 4 dataset. They conducted an aspect-based sentiment analysis approach which consisted on studying specific aspects of products such as food, service, price and ambiance. This research was conducted following three phases; data pre-processing which involved part-of-speech (POS) tagging, feature selection using Chi Square because it has been proven to speed up the computation time in the classification process, and classification of sentiment polarity of aspects using naive Bayes classifier. Based on their evaluation results, the proposed system gave promising output and was able to perform aspect-based sentiment analysis with its highest F-measure of 78.12%.

In (Alomari et al., 2017) claimed that Arabic tweets pose a good opportunity for opinion mining research but they were set back due to lack of sentiment analysis resources or challenges in Arabic language text analysis. Their work included Arabic Jordanian twitter corpus[11] in which the tweets were labelled as positive or as negative. These tweets were analysed using different supervised machine learning approaches. Several experiments were conducted using different TF-IDF weights, stemming and n-grams. This initiative showed that SVM classifier with TF-IDF and bi-grams feature was better as compared to naive Bayes classifier. Following an experimental study they finally concluded that SVM classifier using a combination of TF-IDF with stemming and bi-grams showed 88.72% accuracy and 88.27% f-score, their model performed better than other Arabic sentiment analysis research results.

(Ankit et al., 2018) proposed an ensemble classifier that combines commonly used classifiers such as naive Bayes, random forest, SVM and logistic regression to form a single classifier. Their proposal aimed on improving the performance and accuracy of sentiment classification techniques. The proposed architecture included four modules; 1. Data pre-processing module: for pre-processing the data; 2. Feature representation module: for feature extraction from pre-processed tweets, BoW technique was used for converting tweets into numeric representation; 3. Sentiment classification using base classifiers: in which different base classifiers were implemented for sentiment analysis and finally; 4. Sentiment classification using the proposed ensemble classifier (Nahili et al., 2020). The implementations were done in Python. The results showed that the proposed ensemble classifier performed better than standalone classifiers and majority voting ensemble classifier. In addition, as part of their

[11]https://github.com/komari6/Arabic-twitter-corpus-AJGT

study the role of data pre-processing and feature representation in sentiment classification technique was also explored.

2.10.2 Sentiment analysis of reviews (product/movie)

Adapting text reviews publicly available on the Internet is a trending opportunity in the field of sentiment analysis, which is the process of exploring these product reviews to identify the overall sentiment or opinions about a specific product. Online reviews constitute the notion referred to as user-generated content (Chapter 1), and this is of growing attention and a gold mine for researchers, marketing teams and people who might be interested in sentiments, opinions, views and public mood. The tremendous volume of reviews on the web displays the present scheme (manifestation) of user's feedback (reaction). The task of analysing the sentiment present in a given review is a challenging problem due to several reasons. One matter is the subjectivity detection explained in (section 2.8) in the text and the necessity to distinguish the opinionated sentences from the non-opinionated ones. An additional matter that makes it difficult to classify reviews is that in some cases the writer writes many sentences manifesting the same sentiment (e.g positive), and then concludes with one negative sentence that shifts the overall meaning of the entire text. Consequently, this is why there should be better techniques for feature selection. Many studies have explored different methods in classifying the sentiment of product review, some of which implemented machine learning approaches, some applied lexical methods, and some used linguistic approaches or a combination of various techniques. Due to the fact that a lot of reviews data are publicly available and easily accessed with platforms like Twitter, Amazon, Yelp, IMDb and rottentomatoes. Several research have been conducted to analyse and predict sentiment orientation of online reviews.

(Wang and Manning, 2012) proved that for text classification different alternatives of machine learning algorithms show large variation in their performance. They also demonstrated how Bayes (NB) was more suitable than Support Vector Machines (SVM) for small part in sentiment tasks in addition to bi-gram results that showed constant improvements in tasks. They also, proposed a new SVM variant which showed better consistent results on datasets and resulting to this information they demonstrated NB and SVM variants. This paper resulted in several conclusions: Multinomial Naive Bayes (MNB) was a better choice for sentiment analysis tasks;

SVM was more preferred on long reviews; the performance of bi-grams depends on the sentiment tasks; They also concluded that NB-SVM generated the best results and Bernoulli Naive Bayes (BNB) produced poor results compared to MNB.

(Nguyen et al., 2014) implemented a new feature type to check its contribution in document level sentiment analysis. They achieved better results than (Maas et al., 2011) in terms of accuracy with a score of 91.6% on the dataset collected by (Pang and Lee, 2004). They also applied sentiment analysis on a dataset containing 233,600 reviews and their proposal achieved 93.24% accuracy. In this study, a sentiment polarity classification system has been implemented. Primarily, they described a rating feature used to train the model from an external data set of 233,600 movie reviews. Afterwards, the reported results of their classifier and rating feature showed an accuracy of 91.6% and 89.87% respectively on the datasets from different domains.

In (Mesnil et al., 2014) sentiment analysis was used to perform natural language processing aiming to detect the polarity of text document. First, they followed a binary classification problem using IMDb [12] dataset where only positive and negative sentiments were considered. Various machine learning techniques were implemented for this problem. First one based on language models, second one based on consecutive models of sentences and the last one based on the BoW (Bag of Words) model. This work helped in determining how to adapt different models for sentiment analysis. They included a code which is available publicly at http://github.com/mesnilgr/iclr15

In (Tripathy et al., 2016) presented that the reviews and blog datasets obtained from the social networking sites were unsystematic and need classification for a meaningful information. They used various supervised machine learning methods to classify reviews as positive, negative and neutral. In this particular work, they introduced four different machine learning algorithms: NB (Naive Bayes), ME (maximum entropy), SGD (stochastic gradient descent) and SVM (support vector machine) for sentiment classification. They used precision, recall, F-measure, and accuracy as evaluation metrics for the proposed models. This paper helped in classifying movie reviews using supervised machine learning algorithms which was further applied on IMDb dataset using n-gram approach. The results of this paper gave several insights; they concluded that the classification accuracy decreases as the value of n increases in the n-gram approach; better accuracy was obtained when TF-IDF and count vectorizer techniques were combined. And last, hybrid machine learning techniques were considered for better accuracy.

[12]https://www.imdb.com/

(Tiwari et al., 2017) analysed online audits and film ratings using a content based sentiment analysis approach. They considered different supervised machine learning algorithms to classify these reviews. For conclusions three different machine learning algorithms were considered; SVM, ME, NB and these were based on evaluation parameters such as accuracy, recall, F-measure and precision. In this paper to classify film reviews from the rottentomatoes dataset, the authors used the n-gram method where different machine learning techniques have been suggested. This research main conclusion was that in comparison with other studies their results obtained the best accuracy.

In the work of (Miedema, 2018) the goal was to find out why RNN and LSTM models work well for sentiment analysis and how these models work. Their research was based on the principle of compositionality, which states that the meaning of a longer expression depends on the meaning of its predecessors. RNNs were used to perform sentiment analysis because they allow the network to have a memory. Since the author dealt with sequenced text data, having a memory in a network is useful because the meaning of a word depends on the context of the previous text. The main drawback of the RNN is that its capacity of only dealing with short-term dependencies. To solve this problem they proposed a combination of both RNNs and LSTM. The model was trained using word embeddings as feature. It was very sensitive to overfitting, so the model was stopped from training after the fifth epoch. The final model resulted in an accuracy score of 86.74

So in (Law et al., 2017) focused their study on the domain of under performance in large home appliances precisely dishwashers. They constructed two domain specific lexicons related to dishwasher defects (sparkle and smoke). Their work was destined to improving the quality of dishwasher appliances. The authors had taken on different experiments to capture the defects in the products. They used Afinn lexicon in the first experiment to detect the defects but they reported that the other sentiment analysis techniques gave better performance than uni-gram, bi-grams and tri-grams. From the second experiment they came with the conclusion that in discovering the defects logistic regression, neural network and decision tree classifiers performed better. The best reported results were obtained by neural networks and the negative reviews had unsuitable outcome on sales, brand reputation and company profits.

The approach implemented in the work of (Malik et al., 2018) is a modification of the approach stated in (Haider, 2012). For their extended study, 100 reviews about the product 'Fit-Bit' were crawled from the well-known e-commerce Amazon [13] website. They proposed an ontology based sentiment classification framework in which they used various attributes along with their assigned polarity scores. To capture users' preferences regarding different product aspects and attributes they used the formula defined in (Yaakub et al., 2012). Their experimental results showed that the proposed ontology model works effectively. The result obtained proved that when the buyer's choice is the most specific the more the decision making process is accurate.

After an extensive study of the above surveyed work on sentiment analysis on reviews, an obvious conclusion is drawn as the platforms delivered by e-commerce are evolving at an exponential rate, access to information is becoming pretty straightforward. In consequence, growing number of consumers (users) will seek product information from online consumers feedbacks, instead of the information communicated by the product manufacturers. Reviews shared by users are examples of such type of information and they have by now become a key part of customer's purchasing decision process. Hence, a large number of people lean towards online shopping due to the fact that these platforms provide a transparent system for costumers to make informed decision and feel satisfied with it. Considering the rapid evolution of Internet-distributed computing, up to now, we have the ability to process and analyse tremendous amount of data and predict customer choices and future interests. Therefore, it is becoming progressively compelling to quickly and effectively capture users' sentimental orientations and preferences based on online text reviews and comments. In this thesis, movie and product reviews are investigated, using online reviews data from different sources; first one is tweets and the second one is movie reviews using the famous dataset for sentiment analysis IMDb. The problem is tackled using two approaches; for analysing Twitter data a lexicon-based approach is proposed; for predicting the sentiment of movie reviews (IMDb), a supervised learning methodology is implemented using a novel convolutional neural network. Both approaches will be presented in detail in the next chapter.

[13] https://www.amazon.com/

2.11 Comparison of semantic and supervised learning approaches

After a thorough review and comparison of the surveyed literature on sentiment analysis some interesting observation was generated: for one thing the semantic orientation approach to sentiment analysis is unsupervised learning, because it does not require data for training. But, the supervised learning approach is considered as a supervised learning method, which is dependent on training data, the text features and chosen algorithms identify the polarity of a given text (Liu, 2012, Turney and Littman, 2003). In conclusion, the semantic orientation approaches are typically effective and need very little training, still the performances are usually lower than those of supervised learning techniques. Nonetheless, the supervised learning approach is very time-consuming because the model built by machine learning or deep learning classifiers is extremely dependent on both the size and quality of training data (Zhou and Chaovalit, 2008). A couple of research have been conducted to compare the performance of both methods. (Chaovalit and Zhou, 2005) followed both approaches to perform sentiment analysis by using a movie reviews dataset and realised the first comparison in the same domain. They used the scoring approach from (Turney, 2002) for their semantic orientation approach whereas in their machine learning model, they experimented with n-grams. Their results showed that the semantic orientation approach yielded 77% accuracy, and the machine learning model achieved 85.5% accuracy. However, no results reporting their performance comparison of the two approaches on specifically the same dataset were mentioned. Also, in our opinion the data size was too small and we found obvious inconsistency in the number of positive and negative reviews: 285 positive reviews and 47 negative reviews.

Despite the fact that there is broad agreement that supervised learning approaches achieve better performance than the semantic orientation approach, it is essential to specify that studies also point out that the sentiment classifiers built via supervised learning techniques might perform really well in the domain they are trained on, but the performance can drop tremendously, if the same sentiment classifier is tested in different domains. Because the performance of supervised learning approaches depend on the quality and size of the training data (Brooke, 2009, Zhou and Chaovalit, 2008). Opposed to, semantic orientation methods have the ability to identify the words with assigned polarity usually independent of context and hence offer a much broader scope of application than supervised learning techniques (Liu, 2012). Additionally, in some cases the semantic orientation approach might be a better choice than a supervised learning classifier while handling negation. For example, in some proposed work using the semantic approach, negation is commonly dealt with by calculating the polarity scores of the terms and then determining a combined polarity score (Taboada et al., 2011, Choi and Cardie, 2008). Consequently, the semantic orientation approaches are simply appropriate to determining both the polarity and the strength of orientation considering that their result is a numeric score representing an averaged value of the scores from opinionated words in the text. However, it is difficult to catch the sentiment strength using machine learning approach. Due to the training data often consists of positive and negative training datasets, sentiment classifiers built by supervised learning algorithms frequently output binary results. In conclusion, capturing the rating intervals of polarity using the sentiment classifiers does not come easily in the supervised learning approaches (Brooke, 2009, Pang and Lee, 2005).

Each approach has its advantages and drawbacks. Previous research showed both of semantic and supervised learning techniques for sentiment analysis have been implemented in different domains. Table 2.4 shows a brief detailed summary of the surveyed previous research including the approach and features that were used along with datasets, level of analysis and the highest achieved results in terms of accuracy and F1 score:

Authors/year	**Approach/features**	**Dataset**	**Best results**
Fenna Miedma (2018)	SL with word embeddings	IMDb	86.74%
Malik et al. (2018)	LB with attribute scores	Amazon	
Ankit et al. (2018)	SL with BoW	Sent140 HCR, GOP	75.81%-76.85%
Mohamed et al. (2017)	SL with POS	SemEval14	78.12%
Houshmand (2017)	SL	STT	40.5%-46.4%
Alomari et al. (2017)	SL with TF-IDF,n-grams	AJGT	84.73%-88.72%

Law et al. (2017)	SL with 2 lexicons AFINN,n-grams	Dishwasher reviews	
Tiwari et al. (2017)	SL with n-grams,TF-IDF	Rottentomatoes	87.53%-89.64%
Khan and Bashir (2016)	LB with adj only	Movie reviews	76%
Tripathy et al. (2016)	SL with n-grams,tf-idf	IMDb	70%-86.23%
Saif et al. (2015)	LB	Twitter	87.50%
Severyn et al. (2015)	SL	Twitter	
Park et al. (2015)	LB and SL	Amazon	
Ouayang et al. (2015)	SL with Word2Vec	Rottentomatoes	45.4%
Yih et al. (2014)	SL		54%
Kalchbrenner et al. (2014)	SL	TREC	48.5%-86.8%
Dhande and Patnaik (2014)	SL	Movie reviews	80.6%
Nguyen et al. (2014)	SL with rat-feature,n-grams	PL4 IMDb	80.6%
Khanaferov et al. (2014)	SL	Tweets	
Mesnil et al. (2014)	SL with n-grams,TF-IDF	IMDb	86.5%-92.57%
Ortega et al. (2013)	LB	SemEval13	50%
Balage and Pardo (2013)	LB with adj	Self-collected reviews	69%
Ostrawski et al. (2013)	SL		

Fei et al. (2012)	DB	Dictionary	68.9%
Montejo-Ráez et al. (2012)	LB	Self-collected Tweets	63%
Lin et al. (2012)	SL	Amazon	88%
A. Abrahams et al. (2012)	SL	USDT	
Wang et al. (2012)	SL with n-grams,BoW	RT,Amazon, Subj, IMDb	79%-93.6%
Xu et al. (2011)	SL with ling feature	Self-collected reviews	61%
Taboada et al. (2011)	LB	Movie reviews	
Kouloumpis et al. (2011)	LB with n-grams,POS	Edinburgh,EMOT iSieve	75%
Li and Liu (2010)	SL with TF-IDF	Movie reviews	74.7%
Melville et al. (2009)	SL	Blogposts	91.21%
Zhou and Chaovalit (2008)	Ontology with n-grams	Self-collected reviews	72.2%
Godbole et al. (2007)	LB	News	82.7%-95.7%
Kennedy and Inkpen (2006)	SL	Movie reviews	86.2%
Cui and Mittal (2006)	SL	Froogle	
Gamon (2005)	SL	Amazon	86%
Pang and Lee (2004)	SL	IMDb	86.4%
Turney and Littman (2003)	LB with scores	Novels articles	65.27%
Pang et al. (2002)	SL	IMDb	82.9%

Turney (2002)	LB with scores	Pos neg terms	66%-84%

TABLE 2.4: Summary of the main studies on sentiment analysis.

2.12 Challenges of research in sentiment analysis

Typically, sentiment analysis also referred to as opinion mining is considered a particular case of text classification in a natural language processing task. Although sentiment analysis has a limited number of classes, the process of sentiment classification is more difficult than the traditional topic text classification (Pang and Lee, 2008). Generally, when performing topic text classification, classification depends on using keywords, but according to (Turney, 2002) this does not work well in the case of sentiment analysis. The other challenges faced in sentiment analysis emerge from the nature of this problem. Whereas, we explained in previous sections in some cases the negative sentiment might be expressed in a sentence without using any distinct negative words. Furthermore, there is a subtle line between whether a sentence should be identified as objective or subjective. Extracting the sentiment generator; the one who expresses the sentiment in the text is one of the hardest tasks in sentiment analysis. Additionally, the sentiment analysis task extremely relies on the domain of the data. Also, the words sometimes may convey positive sentiment in a specific domain, at the same time they may express another sentiment polarity in another domain (Pang and Lee, 2008). Last, some other writing patterns such as irony, sarcasm, or negated sentences could lead to more difficulties in the sentiment analysis field.

In pursuance of mining the opinions automatically, two diverse approaches for sentiment analysis have been described in detail in previous sections. The approach of supervised learning has accomplished acceptable accuracy, which to a great degree depends on the quality and size of the training data (Basant and Namita, 2016). This obviously demonstrates that the supervised learning approach is with no doubt domain dependent, which is a key factor that leads to a better performance of sentiment classification in the corresponding domain. Yet, it is difficult to capture the sentiment intensity by implementing the supervised learning approach, because the output is frequently binary; positive or negative. The issue found here resides on the fact that most of real life data available on social media illustrates different strengths of sentiment. Although, the essence of the supervised learning methods, it is not adequate for in-depth sentiment analysis, such as phrase level or aspect level analysis (Khoo et al., 2015) this research is taken on both the semantic and supervised learning orientation technique seeking to inspect a better way for fine-grained sentiment analysis by proposing novel frameworks to analyse social media data.

2.13 Research gaps in sentiment analysis

Over the last few years, the semantic orientation approach for sentiment analysis has been focused and implemented either in academia or in business industry, but it appears it runs into a wall when it comes to performance (Taboada et al., 2011, Khan et al., 2016). A few research gaps are identified in the semantic orientation research. At first, the main task in semantic orientation method is building the sentiment lexicon, which is frequently domain dependent and thus, is limited. The general sentiment dictionary has been globally applied in most research, where sometimes it is an automatic generated lexicon, such as SentiWordNet or manually constructed dictionary, such as Hu and Liu Lexicon (Hu and Liu, 2004, Esuli, 2013, Ghosh and Kar, 2013, Cernian et al., 2015). The principal issue found when using general sentiment dictionary is that the polarities of words rely on the context of use. For example, (Hu and Liu, 2004) have collected a list of English positive and negative sentiment words, which they claim are domain independent. However, majority of words found in that list are unclear without context. Due to the components of natural language, previous research point out that domain dependent approaches can obtain better accuracy (Liu, 2015), but, it is tricky to automatically build domain dependent sentiment lexicons, because that method depend on seed words with clear sentiment orientations such as "good" and "bad" (Kim and Hovy, 2004, Qiu et al., 2009). Therefore, the domain dependent lexicon requires manual effort to select sentiment words. As result of the cost in regards to time and effort, a steady domain dependent sentiment lexicon is impossible. Also, the current sentiment dictionaries mostly contain adjectives, adverbs and verbs. Lastly, they involve limited selection of nouns and phrases that indicate sentiment and no slang language which is largely used on social media platforms. Likewise, the semantic orientation approach is unsuitable for handling contextual information. Both the semantic orientation approach or the supervised learning approach, process text using the bag-of-words model by perceiving text as a sequence of words, which grants no attention to the context (Montejo-Ráez et al., 2014). Hence, it is difficult to manipulate linguistic features, such as negation. Another research gap is that there is small number of methods proposed for sentiment analysis on data containing mixed sentiments. Previous work in the field of sentiment analysis are based on the suggestion that each document or sentence contains single sentiment; positive, negative or neutral, whether it is based on document level or sentence level (Liu, 2012, Zhou and Chaovalit, 2008).

In pursuance of narrowing the research gaps presented above, this research is motivated to explore a new approaches to apply sentiment analysis. Our contribution comes in two folds; the first one is to deal with the notions of negation and natural language processing presented in the research gaps. Recently, many research have been done in the domain of social media analytics precisely sentiment analysis; whether using semantic orientation methods or supervised learning algorithms, most of which concentrate on people's sentiment regarding different topics. The problem in analysing social media unstructured data in this manner is not effective in terms of accuracy and provides a generalised idea. In order to make it more specific, in this thesis, we propose a lexicon-based approach to perform sentiment analysis on tweets since they are a reliable source of information, mainly because people share posts (tweets) about everything, either it is about purchasing products or reviewing them. The previously presented approaches are different from ours: First, we perform sentiment analysis at the phrase level, thus the sentiment polarities assigned at a much accurate, precise level. Second, our approach for polarity assignment is also different since we deal with five classes of sentiment (very positive, positive, very negative, negative and neutral) because it catches the change in sentiment strengths and hence conveys real life scenarios. The second fold is for the limitations in the work done on sentiment analysis of online reviews such as movie reviews from IMDb and rottentomatoes or product reviews from Amazon. We propose a deep learning approach using a convolutional neural network with an embedding layer to perform sentiment analysis. The results proved the model to be effective and achieved satisfactory accuracy. These contributions are proposed in this research in pursuance of slimming the mentioned gaps, which are presented in the next chapter.

2.14 Conclusion

In this chapter, the notion of sentiment analysis and its different levels have been introduced. In particular, a thorough discussion of the semantic orientation and supervised learning approaches for sentiment analysis was provided. After a comparison study of both techniques, the strengths and weaknesses of each approach were also presented, thus many research gaps have been identified. In order to narrow these gaps, 1. A system aiming to perform sentiment analysis is proposed in this thesis by following a lexicon-based approach for sentiment analysis of social media data (Twitter), 2. A deep learning approach, in which a new convolutional neural network model with an embedding is built pursuing to analyse and predict the sentiment of online text precisely movie reviews. Both approaches are presented in the following chapters.

Chapter Three

Proposed lexicon-based approach for sentiment analysis of tweets

3.1 Introduction

In the previous chapter, various research on sentiment analysis (SA) have been described and several research gaps have also been spotted. In order to narrow the gaps and accomplish the objectives of this research, two new sentiment analysis approaches are proposed in this chapter. The first one aims to provide a hybrid approach combining lexicon-based and supervised learning methods to conduct fine-grained sentiment analysis on Twitter data (Nahili and Rezeg, 2018), the second one is a new convolutional neural network model for sentiment analysis of social media text (Nahili et al., 2019). This chapter begins by introducing the specifics of research on Twitter data, followed by the dictionary construction then a detailed description and explanation of the proposed approach for analysing tweets, which seeks to provide an approach to conduct sentiment analysis automatically. Afterwards, the chapter continues with the implementation of a system called TweetEcho along with experimental results.

3.2 The specifics of research on Twitter data

Twitter is a real-time, highly social microblogging service that allows users to post short messages referred to as tweets of 280 characters or less. As opposed to other social networks like Facebook[1], Instagram[2] and LinkedIn[3], where users have a bidirectional connection, Twitter has an asymmetric connection of "friends" and "followers" (Russell, 2011). Twitter is a significant phenomenon due to many reasons; its impressively high number of users (Chapter 2), as well as its use as a marketing device by companies in order to built more targeted marketing campaigns to satisfy customers' needs. This particular microblogging site offers an extensive collection of APIs destined for data analytics (Twitter-Dev, 2020)[4]. As presented in the previous chapter, research on Twitter data analysis has been growing exponentially over the past decade. This interest is due to several reasons:

1. First, social media text is a rich source of public information;
2. Second, information in Twitter is open, has clear number of character and well-documented.
3. Third, it is in real-time where the appearance of messages occurs, figuratively speaking, with the speed of thought.
4. Fourth, this site is used to pass and express opinions by different sectors. Twitter content analysis can help to evaluate changes in moods of many users, to reveal their preferences, likes and dislikes.

Based on the above reasons, microblogging content analysis can help to evaluate changes in moods of many users, to reveal their political preferences, likes and dislikes, their choice in favour of one or another candidate during election campaigns. That is why the development of methodologies destined for Twitter sentiment analysis witnessed growing interest in recent years. Most often, the researchers used sentiment analysis. It can be used for political or sociological researches, for analysis of consumer preferences microblog users, and in other cases.

3.2.1 Sentiment Analysis of Twitter data

The Oxford English dictionary defines sentiment analysis as follows:

[1]https://www.facebook.com/
[2]https://www.instagram.com/
[3]https://www.linkedin.com/
[4]https://developer.twitter.com/

'The process of computationally identifying and categorising opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.'

Sentiment analysis also referred to as opinion mining or text mining is a sub-task from natural language processing (NLP). Successfully performing sentiment analysis caters an important component of understanding the semantics of text. It can be applied to different levels of granularity, as for instance document or sentence level (see Section 2.5). Most of the work on sentiment analysis focuses on product reviews (Medhat et al., 2014). The difference between tweets and typical product reviews is that tweets are restricted in length and contain Twitter specific attributes such as hashtags (indicated by the '#' symbol) and author references (@author) that are not necessarily found in product reviews.

3.3 Proposed lexicon-based approach for Twitter SA

As stated by (Wieringa, 2014), the phases of problem identification and objective definition are considered as fundamental starting points of design science methodology for information systems and software engineering (Wieringa, 2014). As presented in Chapter 1 and 2, the real-world requirements and research gaps of sentiment analysis have been identified based on the preceding research. In order to slim the research gaps and provide an efficient approach to sentiment analysis, the first innovative framework has been proposed in this research (see Figure 3.1).

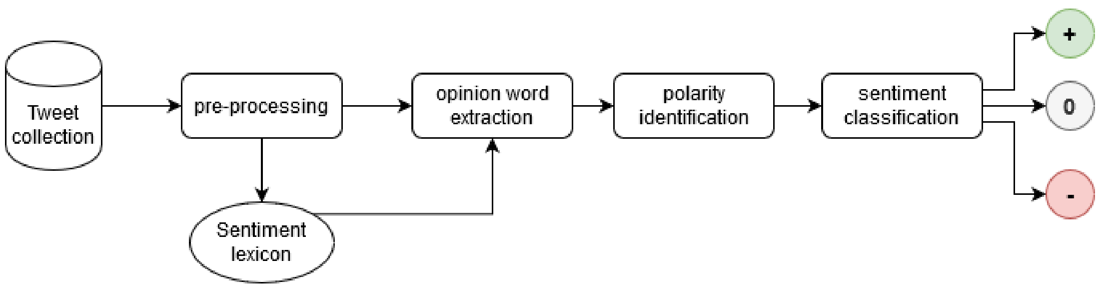


FIGURE 3.1: Proposed approach for Twitter sentiment analysis.

The proposed approach conducts fine-grained (i.e. multi-class) sentiment analysis on tweets at sentence level. It cannot only process and analyse single opinionated text but also mixed opinion text. The system provides an enhanced way to capture phrases or multi-word expressions in the text based on a sentiment lexicon (dictionary) in order to yield more context from the text. Moreover, with the combination of machine learning classifiers specifically naive Bayes classifier the framework offers a valid approach to conduct accurate sentiment classification. The proposed approach is divided into the following four phases: sentiment lexicon construction which will be explained in details in the following sections, data collection, data preprocessing and data classification will be presented is the next chapter.

3.3.1 Sentiment lexicon development

When adapting the lexicon based approach, the main component to conduct sentiment analysis with efficiency is lexicon or dictionary construction. Prior work present the unsupervised approaches to develop dictionary extremely depend on the context of words, but several words have no purpose in the sentiment analysis task, in consequence their unsuitable sentiment orientation values (score) result in a great deal of noise in the final classification of sentiment (Brooke, 2009, Hu and Liu, 2004). With the aim to overcome these problems, the most stable way is manually classifying by human judgment, considering the human annotators are able to identify the polarity of words intuitively in most circumstances and compare words to each other to decide where they should fit on a scale relatively (Taboada et al., 2011, Hatzivassiloglou and McKeown, 1997, Brooke, 2009).

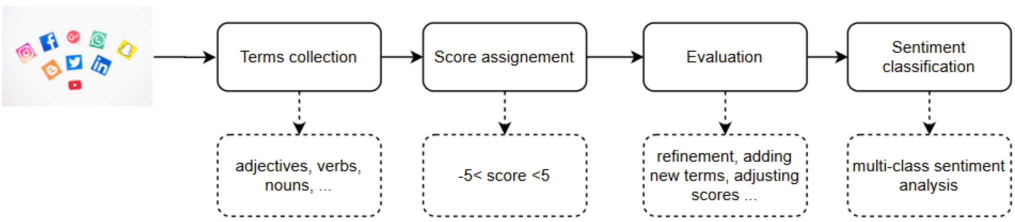


FIGURE 3.2: The different steps for our sentiment lexicon development.

The advantage of this approach is its performance and speed, in addition, it can be used for text analysis at the document, sentence or entity level. In conclusion, the sentiment lexicon within the framework is built manually, which has three steps: collecting frequently used sentiment terms on social media to describe and review products (Nahili and Rezeg, 2018); assigning sentiment orientation value (score); evaluating and refining the sentiment lexicon for necessary modifications further on (Nahili and Rezeg, 2018).

3.3.1.1 Word collection

In this work, we proposed a lexicon-based approach for Twitter sentiment analysis, apart from sentence-level analysis of text, another inventive element of the proposed framework is the development of sentiment lexicon, in which we not only include single-word terms (see Table 3.1) but also multi-word terms (see Appendix A). In this thesis, sentiment terms invoke those sentiment words or expressions used to express positive, negative or neutral sentiments. The sentiment terms in the constructed lexicon can appear in two patterns: single-word term and multi-word term. The single-word term only consists of a single-word entity, whereas multi-word terms contain at least two words such as phrases and abbreviations which are considered as a single entity (Table 3.1). Taking into account sentiment aspects of words in different context, we distinguish two types of sentiment words: global sentiment or domain independent words and domain specific or domain dependent sentiment words. The global sentiment words denotes a selection of words and phrases with polarity consistent across domains such as "good", "love", "wonderful" and "disappointed". What is on the other hand is domain-specific sentiment words, which might only be useful in some particular domain or might have variable polarities in different domains (Turney, 2002, Law et al., 2017, Qiu et al., 2009, Ding and Liu, 2007, Sharma and Bhattacharyya, 2013).

TABLE 3.1: Examples for single-word and multi-word terms from the lexicon.

Single-word terms	**Multi-word terms**
good	no problem
wonderful	best forgotten
satisfied	can’t be happier
bad	not bad
sucks	no good

In this research, the sentiment lexicon is composed of both global and domain specific words by extracting the frequently used sentiment words or expressions in product reviews. There are three reasons why the sentiment lexicon is built over domain specific corpus in our research. First of all, the problem confronted when using lexicon based approach resides on the fact that the polarities of words and expressions rely on the context of application. Previous research show that the domain specific lexicon achieves better sentiment analysis than the domain independent because the expressions of sentiments or opinions differ significantly across different domains. It is challenging to maintain the global lexicon to enclose all the domains (Qiu et al., 2009, Sharma and Bhattacharyya, 2013, Park, 2015b). For example, the word 'soft' might have a positive meaning of warmth in various domains, but it can be negative if it is employed in sport domain (e.g "soft players") (Hamilton et al., 2016). Without selecting the domain specific terms, the sentiment classification task could be mistaken by different contexts.

Secondly, as for the companies or organisations, they are only interested and attentive to the customer reviews or feedbacks associated to their products or services. It is more efficient and useful to develop a domain specific sentiment lexicon for them. Additionally, the domain specific words can simplify the aspect of capturing the explicit and implicit sentiments present in the text. A text with an explicit sentiment conveys a positive or negative sentiment explicitly using a subjective sentence whereas a text that have an implicit sentiment implied the sentiment in an objective sentence (Liu, 2012). For example, the sentence, 'the phone I bought is amazing', expresses an obvious positive opinion. On the other hand, the following sentence, 'the screen lags all the time', illustrates an objective fact but still it implies a negative opinion. The word 'lag' identifies a negative message, hence it can be treated as a specific sentiment term in the phone domain. The sentiment lexicon is developed starting with the collection of a small number of the global sentiment words and phrases, such as 'good', 'bad', which have stable polarity across all domains. Afterwards, the domain specific collection of terms were examined manually and supplementary words and phrases expressing sentiments were collected and added into the sentiment lexicon. For example:

1. (a) *'This is an excellent product.'*

   (b) *'I had an unpleasant customer service experience from this seller.'*

   (c) *'My phone screen is frozen all the time.'*

(d) *'Conserves food frozen well.'*

(e) *'Although this phone looks nice, it lags all the time.'*

The words "excellent" and "unpleasant" in the sentences 1.(a) and 1.(b) respectively demonstrate the positive and negative sentiment, hence they were selected from the selection of words and added into global sentiment terms because they normally convey the same polarity in different domains.

The word "frozen" is commonly indicating negative polarity in a phone review 1.(c), where in the case of a refrigerator review it is positive or neutral 1.(d) Therefore the word "frozen" belongs to the domain specific sentiment terms.

It is frequent to find the word "nice" in a sentence having been used for many positive expressions in various areas, thus it is classified into the universal sentiment terms. The word "lags", for instance, is categorised into the domain specific sentiment terms, since it is commonly used to describe products in some specific domains such as phones and computers.

3.3.2 Score value assignment

The sentiment lexicon developed in this work is established on two principal assumptions: that each word has its polarity, and that the semantic orientation of a particular word can be attributed as a numerical value. A selection of lexicon based approaches have already followed these assumptions (Pang and Lee, 2008, Khan et al., 2016, Nielsen, 2011, Montejo-Ráez et al., 2012). The advantage of affecting the numerical value to a word, is not only able to identify the sentiment orientation but also capture the sentiment intensity. However, assigning semantic orientation scores to a word or phrase can be a difficult task for sentiment analysis (Guerini et al., 2013). A few techniques have been implemented to calculate scores automatically (Esuli and Sebastiani, 2006, Turney, 2002, Kamps et al., 2004) (see Chapter 2). But in light of the absence of stability for automatically generated scores, these approaches have resulted in a large amount of noise in words of calculating the general polarity due to the fact that the automatically generated score commonly does not illustrate the polarity and strength accurately (Brooke, 2009). Therefore, the approach of assigning scores manually is adapted in this research.

In view of various sentiment analysis applications, different researchers have selected different interval ranges for the semantic orientation values (Esuli and Sebastiani, 2006, Taboada et al., 2011). For instance, (Taboada et al., 2011) constructed a dictionary that only contain single word terms, by using an interval ranging from -5 for extremely negative to 5 extremely positive, where 0 identifies a neutral orientation. Nonetheless, the sentiment lexicon built in this research is not only containing single word terms but also multi word terms (see Table 3.1). For example, the terms 'good', 'very good' convey different degree of sentiment intensity. The interval range for the sentiment lexicon in this research needs to be illustrative of different terms found on social media to capture the variation of sentiment intensities.

Therefore in this research, each term collected is added into the sentiment lexicon and assigned manually an integer value varying from negative -5 to positive +5 by human judgement. If a term determines a positive sentiment, it is assigned a positive (+) score, whereas if the term conveys a negative sentiment, it is given a negative (-) score. Logically speaking, for those terms that have neutral sentiments, their score are pointed as zero. The absolute value score of each term expresses the strength of its sentiment polarity (see Table 3.2 below). Still, the use of score interval range from -5 to 5 in the sentiment lexicon can still be somewhat arbitrary.

TABLE 3.2: Examples of adjectives and verbs in the sentiment lexicon.

Terms	**Score**	**Verbs**	**Score**
beautiful	4	like	3
good	3	hate	-3
bad	-3	enjoy	2
really good	4	love	4
disappointed	-3	work	1
extremely beautiful	5	crash	-2
awful	-4	froze	-2
happy	3	loose	-3
sad	-3	recommend	3

3.3.3 Acquisition of single word terms for the lexicon

The manually labelled sentiment dictionary in this research, does not only consist of adjectives (Table 3.2) and adverbs (Table 3.2) that have been proven to be indicators of sentiment in prior studies (Taboada et al., 2011, Turney and Littman, 2003), but also nouns and verbs. Considering the fact that nouns and verbs have different forms, thus they are more complex than the adjectives and adverbs. For instance, majority of nouns have singular and plural structures. As for verbs, they have different conjugation tenses such as past tense, present tense and future tense. In order to be able to analyse the given text efficiently, diverse schemes of nouns and verbs are also collected and added into the sentiment lexicon and they are treated as single words by the proposed system (see Table 3.3).

TABLE 3.3: Examples of adverbs and nouns in the sentiment lexicon.

Adverbs	**Score**	**Nouns**	**Score**
accidentally	-2	masterpiece	5
astoundingly	3	complaint	-3
beautifully	3	disaster	-5
desperately	-3	happiness	3
badly	-4	failure	-2
outstandingly	5	damage	-3
sadly	-3	frustration	-2
perfectly	5	disappointment	-3

A number of conversational terms, common misspellings, "text or chat speak" and informal language frequently witnessed in use in social media are also handled in single word terms. These would include words such as ’luv’, ’gud’, ’lol’, ’lmfao’ and ’haha’ for more examples see Appendix A.

3.3.4 Acquisition of multi-word terms for the lexicon

In conjunction with collecting single word terms, we also considered multi-word terms, such as phrases expressing sentiment orientation were also collected from social media and added into our sentiment lexicon, by cause of multi-word expressions can offer more information to determine the polarity of the text. Processing and analysing multi-word expressions is always challenging for natural language processing tasks (see Chapter 2). But, the proposed approach in this work has the ability to handle multi-word terms directly. Therefore, linguistic features, such as intensification, negation can be treated effectively.

3.3.5 Handling intensifiers

The intensifier is a linguistic term and a modifier, its role is to enhance emotional context to the neighbouring word. The intensifiers are usually used in spoken and written language, since according to (Xiao and Tao, 2007) the degree expressed is of *'a subjective nature, as it reflects and indexes the attitude of the speaker or writer'*. There are two kinds of intensifiers: amplifiers and downtoners. On one hand amplifiers increase the semantic intensity of a neighbouring word, while on the other hand downtoners decrease it (Brooke, 2009). For example, the words such as 'very', 'absolutely', 'completely' are considered as amplifiers, while the words such as 'barely', 'hardly', 'slightly' diminish the force of other words and are referred to as downtoners. More precisely, intensifiers do not convey any sentiment orientations, but they affect the semantic intensity of the words they alter. Thus, the intensifiers and the words altered by intensifiers were collected and added together into the sentiment lexicon as one single new component. For instance, 'this is a very bad phone', the expression 'very bad' present in this sentence has identified strong negative emphasis, hence 'very bad' was added into the sentiment lexicon and assigned the score value of -5, while the term 'bad' has the score value of -4. Table 3.4 illustrates some examples of intensified phases.

3.3.6 Handling negation

Dealing with the negation aspect is an important feature and a challenging task in sentiment analysis considering that, the sentiment orientation of text to be analysed can be entirely reversed by negation terms such as 'not', 'never' and 'barely'. In this work, expressions of negation were also collected and added into the sentiment lexicon, so the system can deal with them directly. For example, in the sentence 'the battery is not bad', here the key sentiment word is 'bad'. Despite the fact that the word 'not' by itself does not express any sentiment orientation whatsoever, it has modified the sentiment of this sentence when it is used with 'bad' together as a phrase. Therefore the term 'not bad' was added into the sentiment lexicon. For another example, let us consider the following sentence 'I don't like this phone', here the expression 'don't like' expresses the polarity of this entire sentence, thus it is also included in the created sentiment lexicon. There are other examples of terms presented in Table 3.4 below:

TABLE 3.4: Examples of negation and intensifiers in the sentiment lexicon.

Term	Score	Term	Score
bad	-4	not good	-4
very bad	-5	can't stand	-3
usable	2	does not work	-3
barely usable	-4	don't like	-3
love	4	dislike	-3
really love	5	no good	-4
extremely love	5	don't love	-4

3.3.7 Evaluation of sentiment lexicon

In this proposed framework for analysing sentiment orientation in social media text specifically tweets, the performance of sentiment analysis principally depends on construction of the sentiment lexicon (dictionary). Accordingly, the process of evaluating and refining the sentiment lexicon plays a major role in the proposed system, thus has a key impact on its performance.

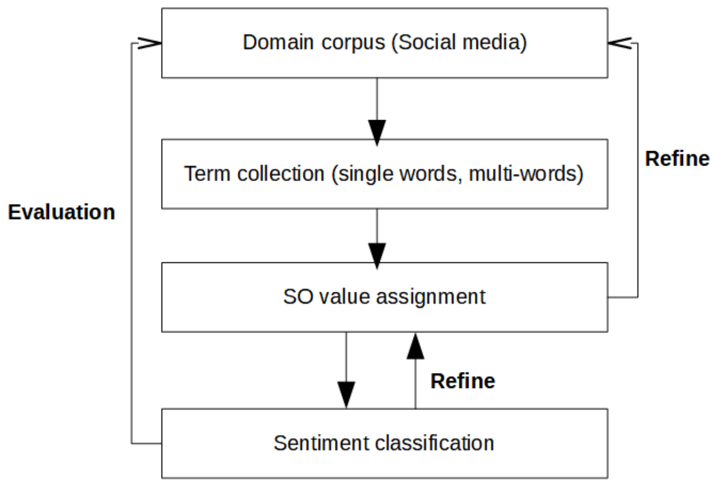


FIGURE 3.3: Different phases for our sentiment lexicon creation.

The task of building the sentiment lexicon itself is literally the process of evaluation and refining in essence. During the collection and selection of sentiment terms into the lexicon, more domain corpora specific to social media are analysed and investigated by updating sentiment terms at the same time. In the course of this process, hence the sentiment lexicon will be refined by adding new words or modifying the score values (see Figure 3.3). The performance of the entire proposed system relies upon the quality of the sentiment lexicon which makes refining the lexicon crucial.

Furthermore, in addition to evaluating the sentiment lexicon while building it, the other evaluation method used in this research is by the comparison of the created dictionary with prior built sentiment lexicons by the judgment of different human annotators (Taboada et al., 2011, Chamlertwat et al., 2012). There is a consistent difficulty in assigning fine-grained score values to the words when it comes to constructing a sentiment lexicon, exclusively when they are out of context. In spite of the fact that the multi-word terms offer additional information to improve the performance, the score value ranges in the fixed interval of -5 to 5 can be very subjective. In this research, the sentiment terms were collected and assigned the score values by the author of the thesis. In order to reduce the bias, the most efficient way to validate the score value of sentiment terms is to compare it with different human annotators to assign the values (Taboada et al., 2011, Chamlertwat et al., 2012). In this thesis, a comprehensive comparative study was followed in order to assign the score values during the process of validating the sentiment lexicon.

3.3.8 Text segmentation

Most of the previously proposed works primarily focus on document level sentiment analysis, which designates the overall sentiment orientation of a given document (see Chapter 2). This issue in analysing sentiment at this level is the fact that the systems assume that each document only contains one single opinion regarding one entity. Having said that, it is frequent that a document might enclose different sentiments in the real world. Therefore, it is inadequate to simply classify the document into one of the polarity categories (positive, negative or neutral) because it does not represent reality. Although more recent research have been conducted at sentence level sentiment analysis (Jain and Pandey, 2013, Alomari et al., 2017), their approaches are also based on the suggestion that a single sentence only contains a one single sentiment. Essentially, there is no major difference between document level and sentence level sentiment analysis (see Chapter 2) in the existing literature.

Consequently, it is difficult to deal with composite and complex sentences that express various sentiments. To illustrate with the following example: "the camera of this new phone is amazing but I am very disappointed at its battery that only lasts for two hours a day". This customer review expresses both positive and negative sentiments (or it has mixed sentiments) regarding the same product, which makes the sentiment classification process even more challenging. To solve these issues, the proposed approach in this thesis implements a novel way to conduct text segmentation at clause level.

The proposed approach is able to separate the text (tweet) into different sections or pieces by tokenisation (see Chapter 4). Taking the following customer review for example, 'I have to say the camera is amazing but I am not happy with its price. Poor signal at times', this customer review about mobile phone expresses positive and negative sentiments regarding different features of the phone. According to the proposed approach, this review will be divided into three sections, as result, the review can be processed and analysed more accurately in the form of three pieces, that is:

(1) 'I have to say the camera is amazing';

(2) 'I am not happy with its price';

(3) 'Poor signal at times.'

3.3.9 Polarity calculation

As described in the previous chapter (section 2.7), the second main task of sentiment analysis is determining sentiment polarity or polarity calculation, which is classifying the polarity of a given text. The majority of the sentiment analysis research based on the semantic orientation approach have primarily investigated document level or sentence level classification (Pang and Lee, 2008, Turney, 2002, Anbananthen and Elyasir, 2013).

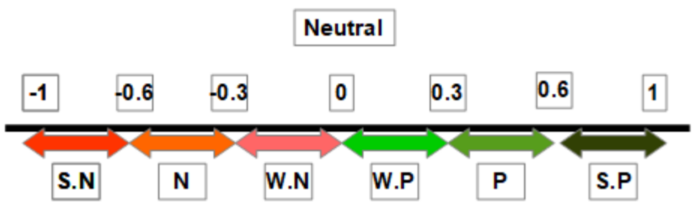


FIGURE 3.4: Predefined intervals for polarity calculation.

In this research, the proposed lexicon based framework for sentiment analysis provides a score based approach to conduct multi-class sentiment classification utilising the manually built sentiment lexicon (see section 3.3.1). Instead of processing the whole document or sentence as one piece or bag of words like other researchers did (Montejo-Ráez et al., 2012, Melville et al., 2009), the proposed system analyses the text at clause level based on text segmentation (see Section 3.3.8). The polarity of each clause or section of text is generated by the values of the sentiment lexicon applied as implemented by the proposed algorithm of the framework.

For our lexicon based approach, we concentrate on the selection of sentiment words frequently used to express sentiment towards an entity (product) on social media. In case of the presence of an irrelevant term to the constructed dictionary, the statement (tweet) is not taken into consideration and thus is rejected. To do so, for each pre-processed tweet we calculate its polarity and assigns a score, after that, we calculate the sum of the polarities to find the average polarity later on.

As illustrated in Figure 3.4 The distinction between positive, negative and neutral statements is calculated as follows: if Sentiment Polarity (SP) is equal to 0 then neutral, if SP is between 0 and 0.3 then weak positive, if SP is between 0.3 and 0.6 then positive, if SP is between 0.6 and 1 then strong positive, if SP is between -0.3 and 0 then weak negative, if SP is between -0.6 and -0.3 then negative and if SP is between -1 and -0.6 then strong negative (Nahili and Rezeg, 2018). For example, as for the customer review of iphone 7 shown below:

'I usually love iphone products but I am very disappointed at iphone 7 this time, poor design and the camera is too bad.'

According to the proposed approach, this review is divided into the following sections at clause level:

1. *'I usually love iphone products;'*
2. *'I am very disappointed at iphone 7 this time;'*
3. *'poor design';*
4. *'the camera is too bad';*

Then the polarity of each segment is determined by making use of the sentiment lexicon and the overall value score can be calculated by summing up of the score values

TABLE 3.5: Example of sentiment classification.

clause	term	score
I usually love iphone products	love	4
I am very disappointed at iphone 7 this time	very disappointed	-4
poor design	poor	-2
the camera is too bad	too bad	-4
Overall score		-6

of the each section, which generates a result indicating that the review expresses negative sentiment or opinion. Adapting this way for sentiment classification offers the following advantages: 1) It captures mixed sentiments; 2) It supports and handles intensification, negation misspellings and slang language;

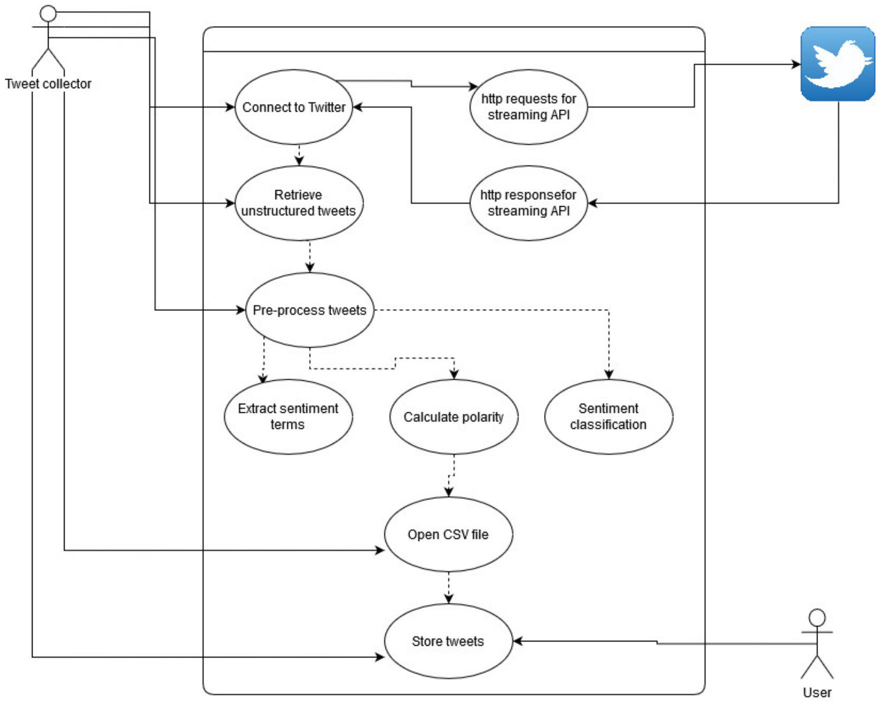


FIGURE 3.5: Use case diagram of the proposed approach for sentiment analysis of tweets.

3.3.10 Naive Bayes classifier

Over these last few years, a lot of research have been done in the field of data analytics precisely sentiment analysis in natural language and social media posts. To determine whether a piece of text expresses a positive or negative sentiment, different approaches are frequently used such as support vector machines (SVM), logistic regression (LR), maximum entropy (ME)...etc. However, most popular classifiers in the field of sentiment analysis are: naive Bayes (Houshmand, 2017, Go et al., 2009, Dhande and Patnaik, 2014), SVM and convolutional neural networks Houshmand (2017), Severyn and Moschitti (2015), Miedema (2018), Lin et al. (2014), thus in the proposed lexicon based approach for sentiment analysis of social media text we combine the constructed lexicon with the machine learning classier naive Bayes for sentiment classification.

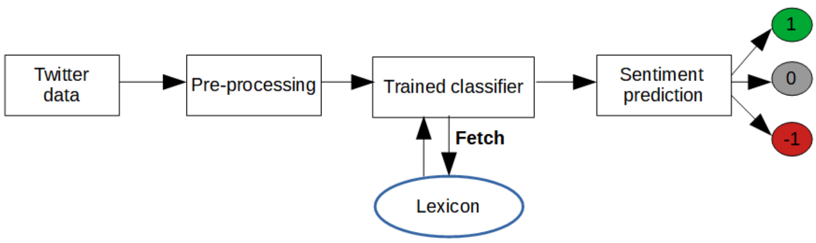


FIGURE 3.6: General workflow of the proposed approach for classifying tweets.

The classifier based on naive Bayes algorithm is a simple probabilistic classifier based on applying Bayes' theorem (equation 3.1), which "applies to a certain class of problems, namely those that can be phrased as associating an object with a discrete category". They are among the simplest Bayesian network models (NB).

$$P(C_k|x) = \frac{P(C_k)P(x|C_k)}{P(x)} \tag{3.1}$$

- P(C|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(C) is the prior probability of class.
- P(x|C) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

After the validation of the created lexicon, the machine learning algorithm naive Bayes is implemented to learn those terms. In this work we experimented with naive Bayes classifier because it is easy to build and particularly functional for large datasets, and it is known to outperform even highly sophisticated classification methods. In view of the fact that naive Bayes is a supervised learning algorithm, it leads to better results when compared to unsupervised methods (Nahili and Rezeg, 2018, Ray, 2015).

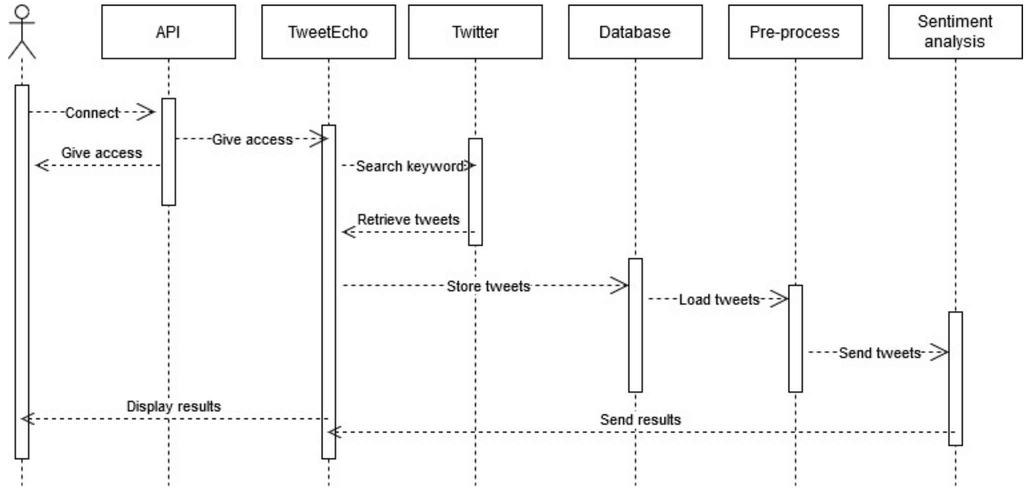


FIGURE 3.7: Sequence diagram of the proposed approach for sentiment analysis of tweets.

3.4 Implementation and Experimental Results

A system called TweetEcho has been established as the physical implementation of the lexicon-based approach presented previously. Our system is able to identify both the sentiment orientation and at the same time its strength from text, utilising a manually developed sentiment lexicon. The proposed system conducts fine-grained (i.e. multi-class) sentiment analysis on tweets at sentence level. It can not only process and analyse single opinionated text but also mixed opinion text. The system provides an enhanced way to capture phrases or multi-word expressions in text based on a sentiment lexicon in order to yield more context from text. The proposed approach shown in Figure 3.8 illustrates the basic inputs, outputs and components of the TweetEcho system.

3.5 Experimental setup

The proposed lexicon-based approach was implemented using the following environment (Table 4.1), we created a setup with the following system requirements.

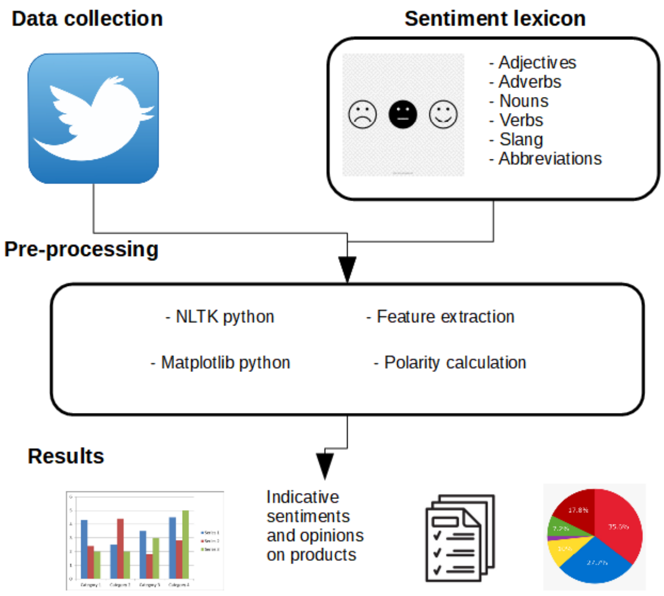


FIGURE 3.8: Inputs, outputs and key components of TweetEcho.

We tested our application 'TweetEcho' (Figure 3.8) on Windows platform. We used a Dell XPS 13 with Windows 10 core-i5 (64 bit) machine equipped with 8GB of RAM. The general requirements shown are in Table 4.1 below.

TABLE 3.6: General system requirements for our lexicon-based approach for SA of Twitter data.

Component	Type
Operating system	Windows 10
Processor	Intelcore-i5 (64 bit)
Memory (RAM)	8GB
Software and Third-party tools	Excel, NLTK, Spyder, Matplotlib

The tools and technology used as follows:

Python 3.6[5] (implementation language): Python is a general purpose, interpreted high level programming language whose design philosophy emphasises code readability. Its syntax is clear and expressive. Python has a large and comprehensive standard library and more than 25 thousand extension modules.

[5]https://www.python.org/

Spyder 3.6.4[6] **(development environment):** is a powerful scientific environment written in Python, for Python, and designed by and for scientists, engineers and data analysts. It offers a unique combination of the advanced editing, analysis, debugging, and profiling functionality of a comprehensive development tool with the data exploration, interactive execution, deep inspection, and beautiful visualization capabilities of a scientific package.

NLTK 3.4.5 [7] **(language processing modules):** The Natural Language Processing Toolkit is an open source language processing module of human language in python. Created in 2001 as a part of computational linguistics course in the Department of Computer and Information Science at the University of Pennslyvania. NLTK provides in-built support for easy-to-use interfaces over 50 lexicon corpora. NLTK was designed with four goals in mind:

1. *Simplicity:* Provide and intuitive framework along with substantial building blocks, giving users a practical knowledge of NLP without getting bogged down in the tedious house-keeping usually associated with processing annotated language data.

2. *Consistency:* Provide a uniform framework with consistent interfaces and data structures, and easily guessable method names.

3. *Extensibility:* Provide a structure into which new software modules can easily by accommodated, including alternative implementations and competing approaches on the same task.

4. *Modularity:* Provide components that can be used independently without needing to understand the rest of the toolkit.

Matplolib 3.1.3[8] **(plotting):** It is a Python 2D plotting library which produces publication quality figures in a variety of hard copy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and four graphical user interface toolkits.

[6] https://www.spyder-ide.org/
[7] https://www.nltk.org/install.html
[8] https://matplotlib.org/

3.6 Data collection

3.6.1 Twitter API

APIs to access Twitter data can be classified into two types based on their design and access method: REST API is based on the REST architecture now popularly used for designing Web APIs. For data retrieval the *pull* strategy is used, in which a user is required to explicitly request the API to collect information. However, the streaming API grants a continuous stream of public information from Twitter. For data retrieval the *push* strategy is used. Once a request for information is granted, the streaming API provides a continuous stream of updates with no need for further input from the user. In view of the fact that in this research, our goal is proposing a real-time Twitter sentiment analysis system we used the streaming API. In order to acquire Twitter data programmatically, we needed to create a Twitter developer application which is used to interact with Twitter's streaming API. This API can be accessed only via authenticated requests. Twitter applies "open authentication" and each request must be identified with valid Twitter user credentials which were generated when the developer application was created following the subsequent steps:

- Create a Twitter account or sign into an existing account;
- Using a Twitter account, sign into the Twitter Developers;
- Navigate to My Applications;
- Create a new application;
- Fill out the new application form;
- Scroll down and click on Create my access token button;
- Record Twitter access keys and tokens;
- Install and load Tweepy python packages;
- Create and Store Twitter Authenticated Credential Object;
- Authorize Twitter application to use the account;
- Extract tweets;

```python
import tweepy
from tweepy import OAuthHandler

consumer_key = 'YOUR-CONSUMER-KEY'
consumer_secret = 'YOUR-CONSUMER-SECRET'
access_token = 'YOUR-ACCESS-TOKEN'
access_secret = 'YOUR-ACCESS-SECRET'

auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_secret)

api = tweepy.API(auth)
```

FIGURE 3.9: Python code to access/fetch twitter API.

Additionally, to access the streaming API and stream tweets, we installed the Python package *Tweepy* due to the fact that it is pretty straightforward to use and the Figure 3.9 below illustrates the python code needed to fetch the API. For our approach, since the developed lexicon is specific to social media text and is destined to help improve the decision making process for organisations, we prioritise a list of different selection of random keywords (terms) to retrieve tweets, these terms must specify an entity that can be a product such as smartphone brands or laptops.

3.7 Data pre-processing

Data pre-processing is a crucial step when performing any sentiment analysis task since with the proper selection of pre-processing methods, the classification accuracy can be enhanced (Haddi, 2015, Parikh and Movassate, 2009). As explained in section 3.2, Twitter data has its own characteristics, thus we apply both Twitter specific and standard pre-processing on the collected data (tweets) via the streaming API (Figure 3.10). The specific pre-processing is primarily important for tweets, due to the fact that Twitter community has developed its own unique phrases and forms to write and share posts (tweets). Furthermore, user-generated content in social media commonly caters slang, in addition to frequent grammatical and spelling mistakes (Petz et al., 2013). Therefore, with Twitter-specific data pre-processing we try to deal with these properties of the Twitter language and improve the performance of our proposed lexicon-based approach. To do so, we consider the following steps for the pre-processing task in our proposed system:

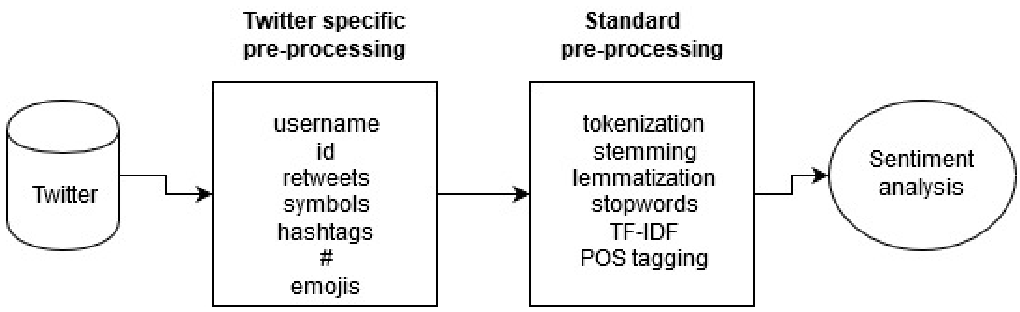


FIGURE 3.10: The different steps for the pre-processing task of our system.

3.7.1 Standard pre-processing

Standard pre-processing involves the succeeding tasks:

1. Remove numeric and empty texts;

2. Remove punctuation from texts;

3. Convert words to lower case;

4. *Tokenisation:* is the process of converting the text sequence of 280 characters composing the tweets, into a sequence of words (tokens), in the proposed framework for Twitter sentiment analysis only the tweets containing matching terms (words) from the predefined selection of sentiment words specific to the social media domain (Chapter 3) are extracted and stored for further processing. In simple terms, tokenisation means dividing a given text into smaller and meaningful elements like words (Nahili and Rezeg, 2018, Andrew and Markus, 2016).

   For example, let us consider the following product review: "The battery of the iPhone is awful and the display as well". After tokenisation is applied, the sentence will provide the following tokens: "The", "battery", "of", "the", "iPhone", "is", "awful", "and", "the", "display", "as", "well" (Nahili and Rezeg, 2018)

5. *Stopwords removal:* Twitter data is by nature noisy seeing that natural language is used, therefore, very basic and rudimentary cleanup was performed using the predefined stopword list from the *nltk* Python package;

6. *Text normalization:* It consists of two techniques: stemming and Lemmatisation, which is an important step in order to get better performance for the proposed lexicon, and it is basically preparing the tweets for further processing.

Similar to stopwords removal, to stem and lemmatise words and sentences, we used the Python *nltk* package which is the Natural Language Toolkit package provided by Python for natural language processing tasks.

Tweet 1	b'Amazing.. We love \xe2\x9d\xa4 Xiaomi devices... #ILoveXiaomi #Xiomi \xe2\x9d\xa4 #RedmiNote8Pro #RedmiK20Pro'
Tweet 2	b"#ILoveXiaomi recently bought it it's truly amazing with lots of features I love xiomi\xe2\x9d\xa4\xef\xb8\x8f\xf0\x9f\x8f\xa9"
Tweet 3	b'No doubt that Everyone loves xiomi My favorite smartphone brand.. \xf0\x9f\x91\x8c\xf0\x9f\x91\x8c #ILoveXiaomi amazing features https://t.co/psLiz6N4VU'
Tweet 4	b'Everyone loves xiomi My favorite smartphone brand.. \xf0\x9f\x91\x8c\xf0\x9f\x91\x8c #ILoveXiaomi amazing features https://t.co/38eisVJXaq'

FIGURE 3.11: Unprocessed tweets samples from accessing the streaming API using 'xiomi' as keyword.

3.7.2 Twitter specific pre-processing

As presented in Chapter 3 a tweet has its own structure, which is composed of different properties such as retweet symbol "RT", emoticons, special characters, hashtags "#" and words that start with "@" character known as username and finally links/URL. But for the purpose of our study, all these components were removed since we found no significance for them in our score based approach, in addition duplicated tweets were also removed. Last, retweeted tweets with no "RT" symbol were also removed. After the pre-processing phase the clean tweets are stored in a comma-separated values 'result.csv' file for further processing.

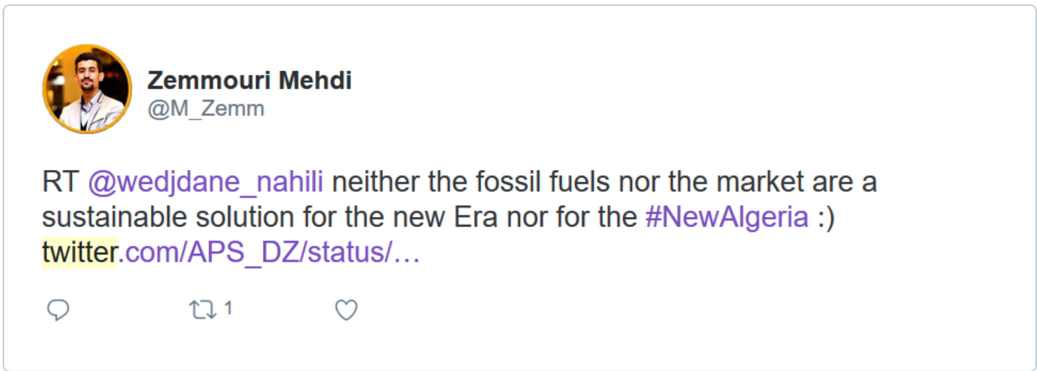


FIGURE 3.12: Tweet example highlighting its different components (features).

Let us take the following tweet with 'Xiomi' as a keyword to fetch Twitter's streaming API. The first tweet shows the tweet before pre-processing and the second one shows the tweet after pre-processing.

Before pre-processing: "b'Amazing.. We love x e2 x9d xa4 Xiaomi devices... #ILoveXiaomi #Xiomi xe2 x9d xa4 #RedmiNote8Pro #RedmiK20Pro"

After pre-processing phase the above tweet example would look like the following: "amazing love **xiaomi** device"

Where: 'amazing' and 'love' as sentiment terms from the lexicon, 'xiaomi' as the keyword to fetch the streaming API and 'device' as a domain dependent term from the lexicon.

3.8 Feature selection

Feature selection is necessary in any NLP task in order to improve accuracy. We carry out our analysis implementing several feature selection algorithms, one using our built dictionary, the others using already implemented features from the *nltk* Python package. Due to the fact that we are analysing tweets the used features were regarding words or sequence of words, therefore, we used tf-idf, POS tagging (Part Of Speech), negation, stemming and stopwords to extract useful features from text.

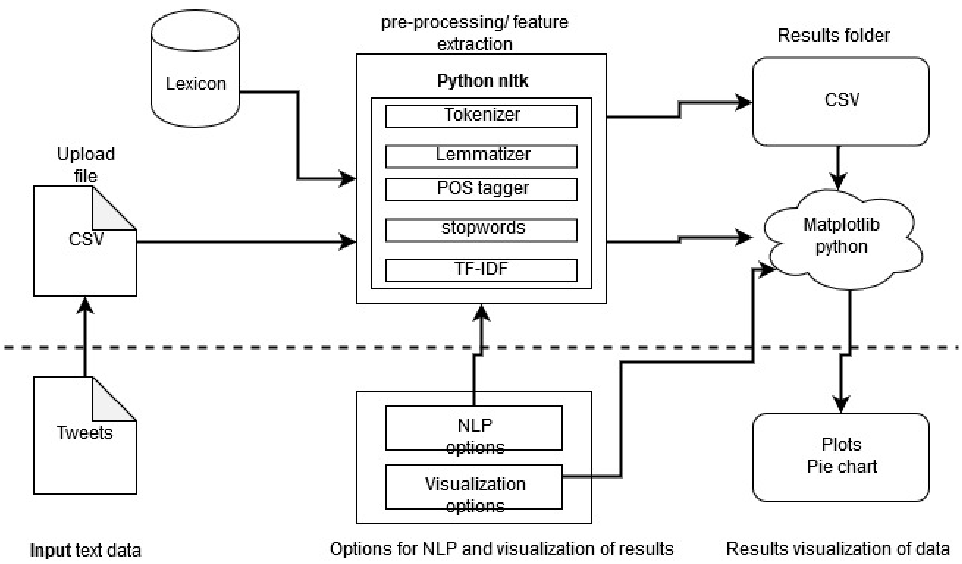


FIGURE 3.13: The different phases of our TweetEcho system.

3.8.1 Part of speech tagging

POS tagging is a linguistic feature that refers to those features incorporating rich linguistic annotation. This key feature is always implemented in most studies (Turney, 2002, Montejo-Ráez et al., 2014, Kouloumpis et al., 2011, Mohamad et al., 2017). Such features usually rely on highly accurate taggers (and parsers). This feature processes a sequence of words, and attaches a part of speech tag to each word. Given the following tweet "Everything is all about politics.", the Part-of-Speech tagging is shown below:

Input: Everything is all about politics.

Output: [('Everything', 'NN'), ('is', 'VBZ'), ('all', 'DT'),('about', 'IN'), ('politics', 'NNS'), ('.', '.')]

Where: **NN:** noun, singular 'desk'

VBZ: verb, 3rd person sing. present takes

DT: determiner

VBZ: verb, 3rd person sing. present takes

IN: preposition/subordinating conjunction

NNS: noun plural 'desks'

3.8.2 TF-IDF (Term Frequency-Inverse Document Frequency)

The TF-IDF feature is by far the most used feature both in natural language processing and sentiment analysis tasks (Pang et al., 2002, Alomari et al., 2017, Tripathy et al., 2016, Wiebe et al., 2004, Li and Liu, 2010). TF-IDF stands for term frequency-inverse document frequency. This feature is a statistical measure used to evaluate how important a word is in a given text (document). The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus (Wu et al., 2008). In simple language, TF-IDF can be defined as follows: A high weight in TF-IDF is reached by a high term frequency (in the given document) and a low document frequency of the term in the entire selection of documents.

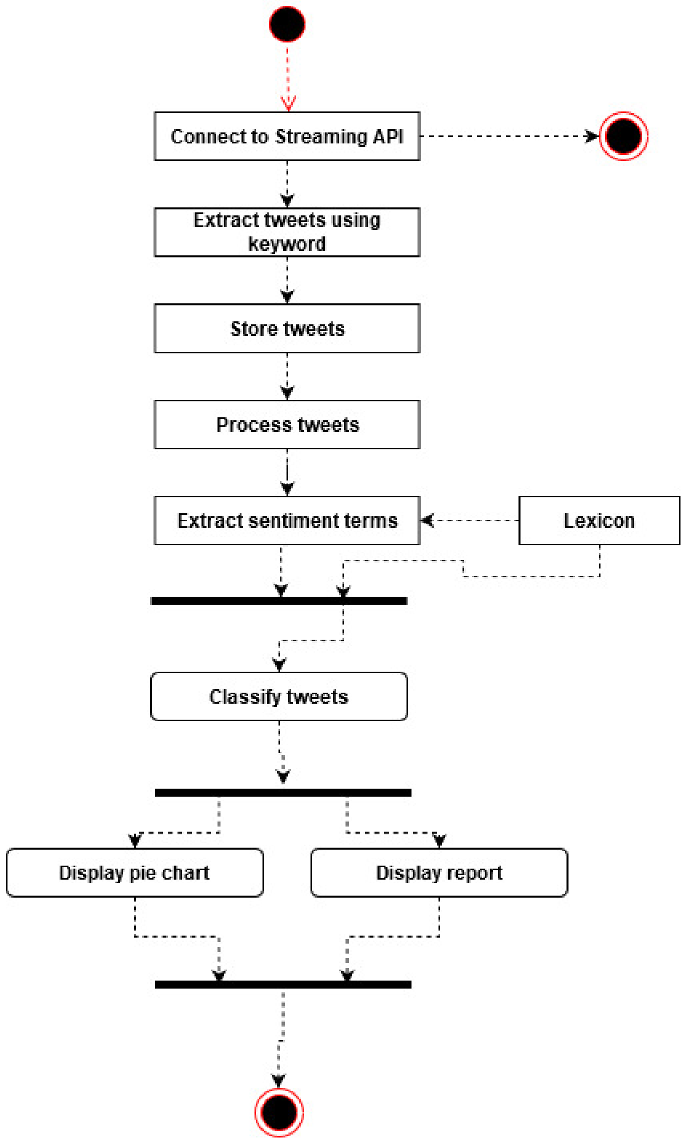


FIGURE 3.14: Activity diagram of the proposed approach for sentiment analysis of tweets.

3.9 Experimental results

The approach can be applied using any product as keyword and several products can be considered for sentiment analysis, for instance "Samsung S10" and the number of tweets to be extracted can be set by the user. In Figure 3.15, we collected 1000 tweets from Twitter via its API. Our proposed lexicon-based framework does not only perform binary-class sentiment analysis (i.e. positive/negative) of tweets but it takes into consideration the intensity of sentiment orientation with our score assignment as explained in Chapter 3, thus, it performs multi-class sentiment analysis (strong positive, positive, weak positive, neutral, weak negative, negative ans strong negative). When performing sentiment analysis using TweetEcho, the system generates two types of output:

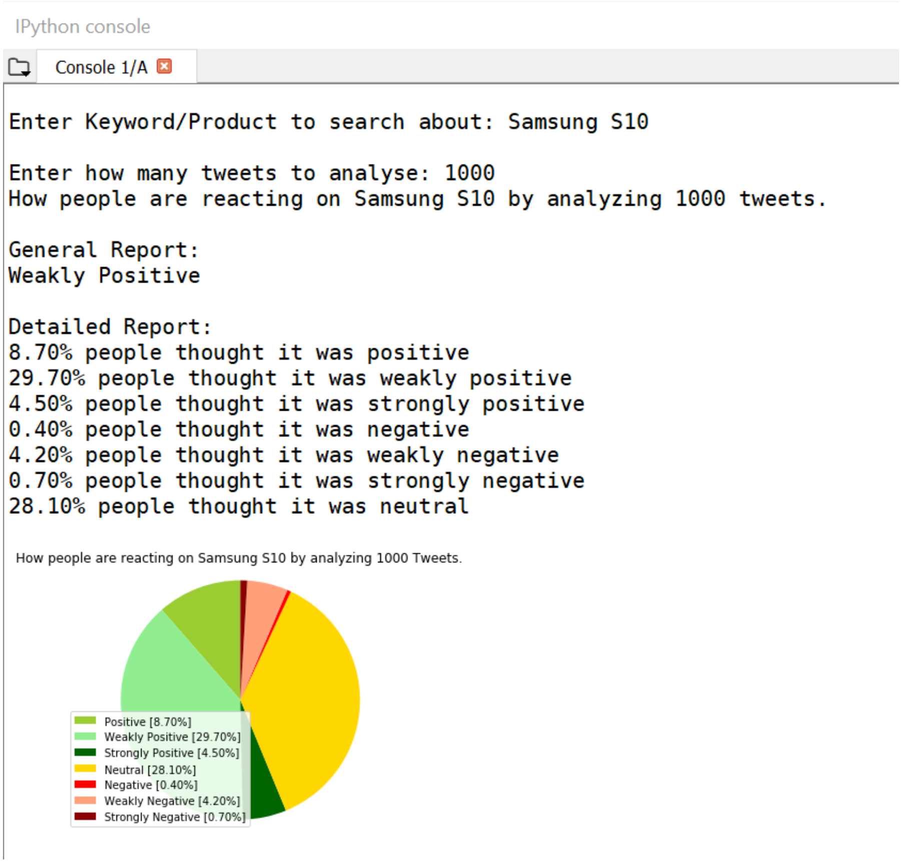


FIGURE 3.15: The output of TweetEcho after analysing the latest 1000 tweets using 'Samsung S10' as keyword (product) to fetch the streaming API.

1) A general report which is the result of the AVG equation (equation 3.2) used to calculate the average reaction (feedback) of users for a specific product (smartphone brand, laptop brand...). To illustrate how our system TweetEcho analyses and classifies tweets we take the smartphone brand 'Samsung' and as an example the keyword 'Samsung S10'. The output of this report can be weakly/strongly positive, weakly/strongly negative or neutral. The average reaction of users is calculated as follows:

$$AVG - polarity = \frac{Polarity}{Numberoftweets} \tag{3.2}$$

- AVG-polarity: The average polarity of users.
- Polarity: The determined polarity after assigning each tweet to its class.

- Number of tweets: The number of collected tweets.

2) A detailed report which is the result of the polarity classification of each extracted tweet to a specific class depending on the polarity interval (Chapter 3 subsection 3.3.9). After determining the percentage of each class, the generated detailed report includes seven sentiment categories including the percentage of each class (see Figure 3.15).

In order to make more sense out of the results, we used the *Matplotlib* Python 2D plotting library which produces publication quality figures in a variety of hard copy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and for graphical user interface toolkits (Nahili and Rezeg, 2018). The below figure displays the output of the proposed TweetEcho system performing sentiment analysis of tweets using 'Samsun S10' as keyword. We used color coordination to illustrate different sentiment categories.

3.10 Conclusion

In this chapter an innovative approach for sentiment analysis of Twitter data was proposed and described in detail. The chapter has also provided the justification and explanation for each component, in order to achieve the aim and objectives of this thesis (see Chapter 1). In order to examine and experiment with this theoretical approach, a sentiment analysis application called TweetEcho was developed as the physical implementation of our lexicon-based approach. Our system is able to determine the sentiment polarity from tweets in addition to its strength. It is also adequate at handling online text aspects such as negation, intensification, mixed opinionated text and slang language primarily from the constructed sentiment lexicon. Although our proposed lexicon-based approach for sentence level sentiment analysis had bridged a few gaps based on the reported results, it still has some drawbacks. For instance, collecting and manipulating Twitter data, additionally evaluating/refining the created lexicon require human effort which is time consuming, therefore, in view of proposing a fully automated classification process we move towards a deep learning approach which will be presented in the next chapter.

Chapter Four

Proposed deep learning approach for sentiment analysis of movie reviews

4.1 Introduction

In the previous chapter, a lexicon-based approach for sentiment analysis of Twitter data was proposed and described thoroughly. Based on the results, several drawbacks have also been spotted. In order to solve the encountered issues and move this research to the next level, a deep learning approach following the supervised learning classification framework is proposed. To do so, a new CNN model with an embedding layer *emb-CNN* is presented in this chapter (Nahili et al., 2019). This chapter begins by introducing a detailed layer architecture of our proposed convolutional neural network specifically designed to predict the sentiment of movie reviews. Afterwards, the chapter continues with the implementation and experimental setup of our model followed by the results along with the discussion.

4.2 Deep learning based approach for text analytics

Convolutional Neural Networks (CNN) are a form of artificial neural networks that can detect information in different positions with excellent accuracy. This model has solved several problems in image processing and automatic natural language processing such as opinion analysis, question answering, text summarization. It is characterized by a particular architecture to facilitate learning. A convolutional neural network is a multilayer network, where the output of one layer will be the input of the next layer. It is usually composed of an input, one to several hidden layers and an output. Recently, the biggest reason to adopt convolutional neural networks in natural language processing, sentiment analysis, text, topic and document classification is due to the following key reasons; - CNN can extract an area of features from global information; - it is able to consider the relationship among these features (Kim, 2014); -Text data features are extracted piece by piece and the relationship among these features, with the consideration of the whole sentence, thus, the sentiment can be understood correctly.

It is stated that convolutional neural networks proved to be successful for classification tasks like sentiment classification since sentiment is usually determined by some key terms and phrases. In this work, we propose emb-CNN a corpus-based convolutional neural network with an embedding layer for text analytics on large scale dataset in order to predict sentiment orientation of movie reviews. Our model's strength characteristics are its efficiency regarding training time and accuracy (Nahili et al., 2019). The proposed model was trained, tested and validated on the publicly available dataset IMDb. In the following section a detailed description of the proposed emb-CNN model is given. Subsequently, sentiment analysis is applied using the model, the results and experimental setup are explained in detail along with the dataset being used for training, testing and validation.

4.2.1 Proposed emb-CNN model

With the aim to automate the process of text classification, several studies have been conducted in different fields like semantic parsing, sentence modeling and sentiment analysis (Ouyang et al., 2015, Houshmand, 2017, Mikolov et al., 2013, Kalchbrenner et al., 2014, Yih et al., 2014). In this work, we propose a convolutional neural network to perform binary-class (i.e. positive or negative) sentiment analysis. In order to do so, we first train the proposed network on data from the IMDb dataset, then we evaluate the model using test data to predict the sentiment of movie reviews. For data representation, we use word2vec as an embedding layer to improve the feature set, hence the model's performance. An overview of the model is shown in Figure 4.1 which represents the process that takes place throughout the sentiment analysis process which is divided into two sub-processes: the learning process which is a typical learning process for any given model, where we train and test and validate the emb-CNN model. Later on, the classification process where basically binary-class sentiment analysis is applied using the trained emb-CNN model. As illustrated below in Figure 4.1; First, the IMDb dataset is collected then before any further analysis of the data, pre-processing is needed, afterwards text vectorization is applied using word2vec as a resource. Last, sentiment classification is conducted using the trained model.

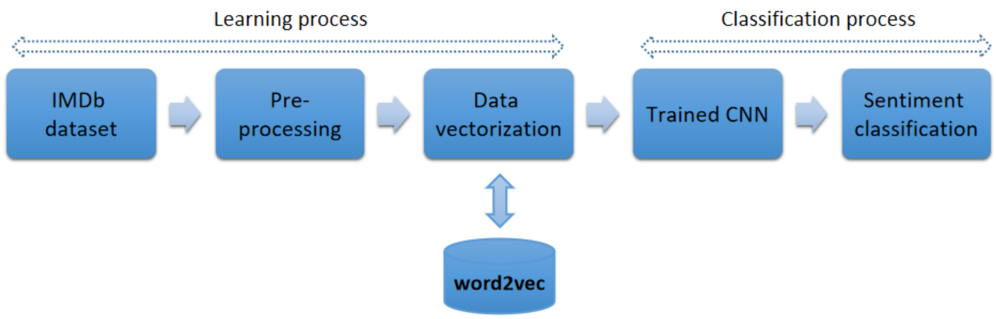


FIGURE 4.1: Global architecture of our deep learning approach for sentiment analysis of movie reviews.

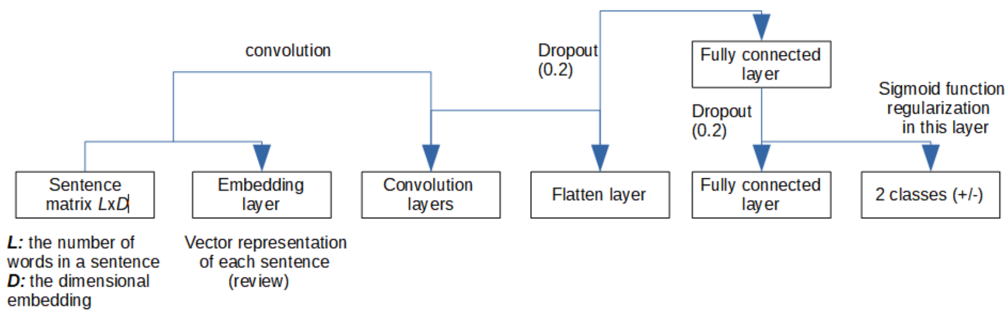


FIGURE 4.2: The layer architecture of the proposed emb-CNN model.

4.2.1.1 Input layer

Instead of image pixels, the input to most natural language processing tasks is sentences or documents represented as a matrix. Each row in the matrix corresponds to one token, typically a word, but it could be a character (Krizhevsky et al., 2012). That is, each row is a vector that represents a word. Typically, vectors in deep learning are basically numerical attributes from which anyone can perform some mathematical operation (Nahili et al., 2019, Gibson and Patterson, 2017). In order to convert string features into numerical features, several methods have been proposed such as Bag-of-Words (BoW), n-grams model and Term Frequency-Inverse Document Frequency (TF-IDF). These are only useful as a lexical resources, but their main drawback resides on the fact that they cannot capture semantics like word embedding. In our work we experiment with word2vec published by Google in 2013, which is a neural network implementation that learns distributed representations for words (Mikolov et al., 2013). Prior to word2vec, other deep or recurrent neural network architectures had been proposed (Ouyang et al., 2015, Kalchbrenner et al., 2014) for learning word representations. The major problem with previous attempts was the long time required to train the models while word2vec learns quickly compared to these models. In order to create meaningful representations word2vec does not need labels. Since most data in the real world is unlabelled, this feature is very useful. If the network is trained on a large dataset, it produces word vectors with interesting characteristics. As result, words with similar meanings appear in clusters, and clusters are spaced such that some word relationships, such as analogies, can be reproduced using vector math (Nahili et al., 2019). For example let us generate a sentence matrix with a 5-dimensional embedding of the following sentence: 'I like this movie very much'. We can see that it is composed of six words as result we would have a 6x5 matrix as input. That is our input sentence matrix (image) to the network.

4.2.1.2 Embedding layer (word2vec)

As input to our proposed model the first layer is an embedding layer which is defined as the first hidden layer and its role is to transforms words into real-valued feature vectors known as embeddings. These vectors are able to capture morphological, syntactic and semantic information about words. It must specify the following arguments: top words appearing in a given review, embedding vector length and the maximum review length. In this work, we truncate the reviews to a maximum length of 1600 words and we only consider the top 10,000 most commonly occurring words in the movie reviews dataset, and we used an embedding vector length of 300 dimensions. This is an important step in the proposed network architecture because it leads to initialise the parameters at a good point of our CNN model.

4.2.1.3 Convolutional layer

The convolution layer aims to explore the combination between the different words/sentences from the reviews, using word2vec. The output of the embedding layer is a 2D vector (None, max-review-length, embedding-vector-length) with one embedding for each word in the input sequence of words (Nahili et al., 2019). Some modification is applied to the basic convolutional operation (layer) where padding is used to conserve the original size of the input sentence matrix, therefore, no loss of information. Since we are working with text data in order to connect the dense layer (fully connected layer) to the 2D output matrix a flatten layer was added to the network in order to convert the output.

4.2.1.4 Flatten layer

The last stage of a convolutional neural network (CNN) is a classifier. It is called a dense layer or fully activated layer. At this stage the classifier needs individual features, just like any other classifier. This means it needs a feature vector. Therefore, we need to convert the output of the convolutional part of the emb-CNN into a 1D feature vector, to be used by the last layer. This operation is called flattening. It gets the output of the convolutional layers, flattens all its structure to create a single long feature vector to be used by the dense layer for the final classification.

4.2.1.5 Regularisation

Regularisation is managed through several functions that organise a complex neural network to avoid overfitting that impacts the performance of deep learning models. With approximately 7 million trainable parameters, the proposed CNN model is very powerful. However, overfitting is a serious problem in large networks making them slow to use and thus difficult to deal with overfitting by combining many different predictions at test time. Dropout is a technique that prevents this problem and it refers to dropping out units (hidden and visible) in a neural network (Lai et al., 2017). By dropping a unit out, we mean temporarily removing it from the network, along with all its incoming and outgoing connections. In our model we use two dropout layers with (0.2), and the choice of which units to drop is random.

4.2.1.6 Fully activated layer (Dense)

In deep learning models, activation functions can be basically divided into two types; linear activation functions and non-linear activation functions (Glossary, 2018). In our work, we experiment with the proposed model to apply binary-class sentiment analysis using IMDb dataset where we used the sigmoid activation function. The main reason why we used sigmoid function is because it exists between (0 to 1). Therefore, it is adequate for our model since we have to predict the probability as an output.

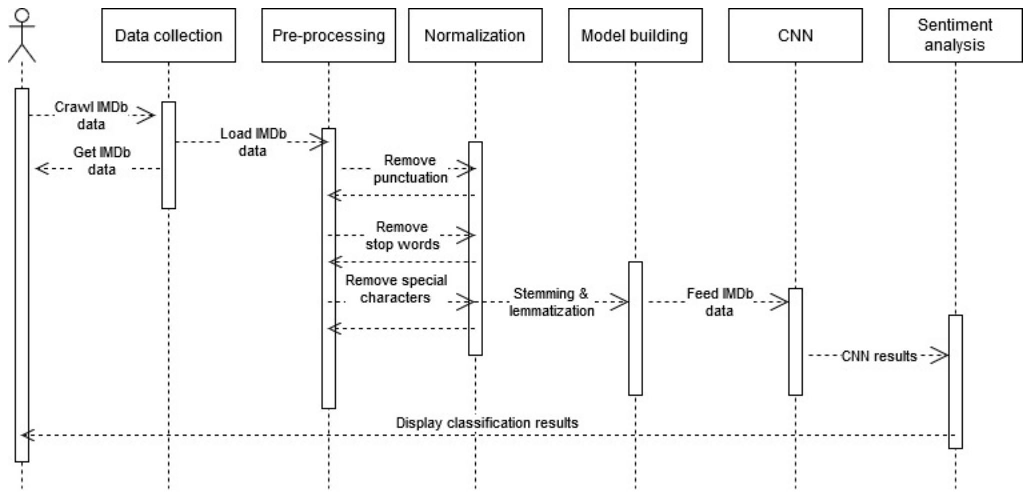


FIGURE 4.3: Sequence diagram of the proposed deep learning approach for sentiment analysis of IMDb reviews.

4.2.1.7 Optimization

Optimization is used in training deep learning algorithms for updating the model parameters (weights and bias values) across iterations. There are different optimization strategies that calculate appropriate and optimum values for these parameters such as Stochastic Gradient Descent (SGD) (LeCun et al., 2012, Bottou, 2010) or Adaptive Moment Estimation (Adam) (Kingma and Ba, 2015). SGD is the classical non-adaptive optimization algorithms used to optimize deep learning networks that use a single learning rate which does not change during training. For optimizing our emb-CNN model we use Adam which is an extension method of SGD that uses an adaptive learning rate to optimize the networks that converges very quickly and outperforms SGD (Kingma and Ba, 2015).

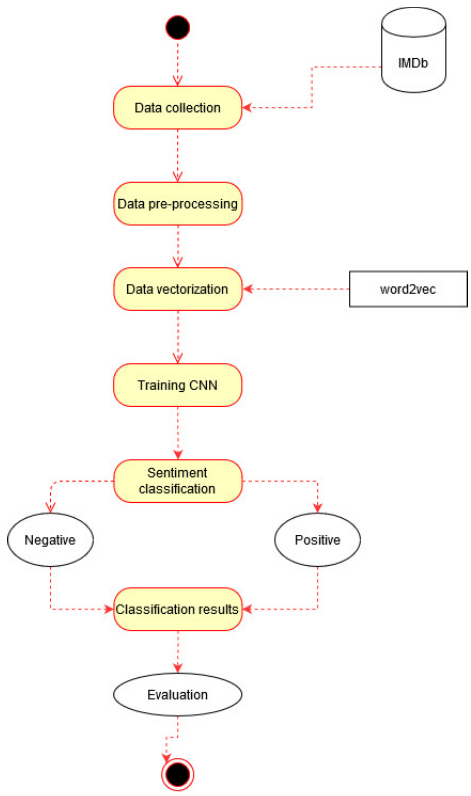


FIGURE 4.4: Activity diagram of the proposed deep learning approach for sentiment analysis of IMDb reviews.

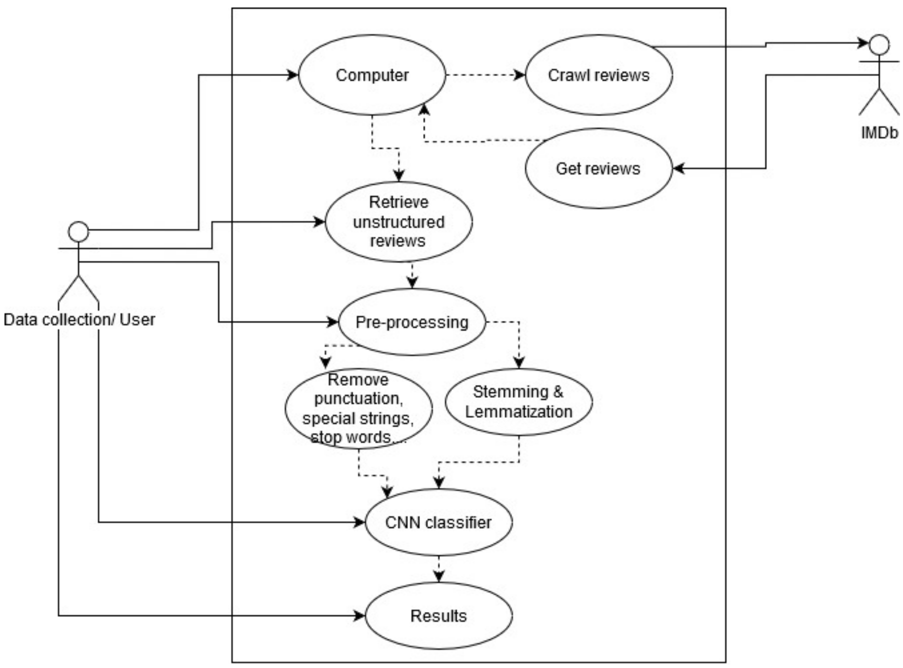


FIGURE 4.5: Use case diagram of the proposed deep learning approach for sentiment analysis of IMDb reviews.

4.3 Implementation and experimental results

4.3.1 Proposed Model: CNN and word2vec for sentence-level SA

In this work, the IMDb dataset is used to illustrate how our proposed emb-CNN model can be used for binary-class sentiment classification. Thus, we propose the following architecture composed of these main three parts, which are described in more detail below in Figure 4.6.

Pre-processing part: In this stage, data cleansing and pre-processing are performed. Then, data representation using word2vec embeddings is applied to prepare the data for the convolution part. The resulting vector is passed as an input to the next stage.

Convolution part: In this stage, convolution layers are applied for feature extraction to extract high level features. The output of this stage is the input of the next stage.

Fully connected part: In this stage, fully connected layers are applied for sentiment classification of IMDb reviews. The output of this stage is the final classification of the review (as positive or negative).

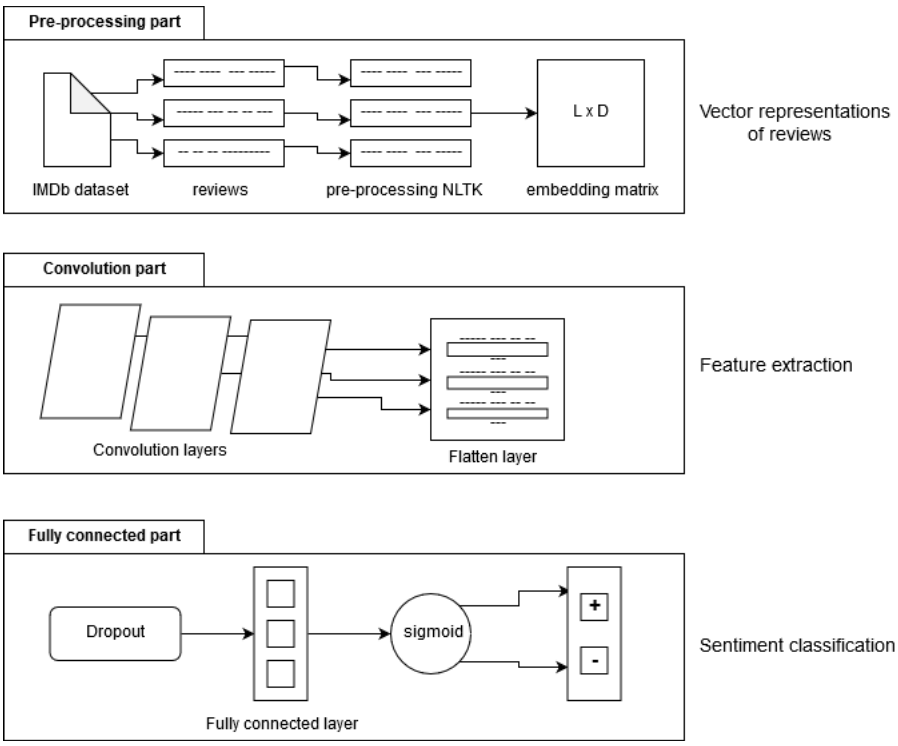


FIGURE 4.6: General architecture of our emb-CNN model.

4.3.2 Data collection

The IMDb dataset is a set of movie reviews from the Internet Movie Database. Each of these movie reviews is classified as either being 'positive' or 'negative'. The dataset was downloaded from IMDb. It is contained in a gzipped, tab-separated-values (TSV) formatted file in the UTF-8 character set. The first line in each file contains headers that describe what is in each column. A "/N" is used to denote that a particular field is missing or null for that title/name. This dataset consists of 50k movie reviews. It provides a set of 25,000 highly polar movie reviews for training, and 25,000 for testing. There is additional unlabelled data for use as well (Maas et al., 2011). The original ratings on the IMDb dataset are 1-10 star ratings, these are linearly mapped to [0,1] to use document labels when training the model. A "1" indicates a positive review and "0" a negative review.

4.3.3 Data pre-processing

Pre-processing the data is the process of cleaning and preparing the text for classification (Haddi, 2015). It is necessary to normalise the text for any natural language processing task. Similar to tweets online text data especially reviews is often represented in a cryptic and informal way, thus contain lots of noise and non-relevant parts such as HTML tags, scripts and advertisements.

	review	**sentiment**
0	One of the other reviewers has mentioned that…	positive
1	A wonderful little production. The…	positive
2	I thought this was a wonderful way to spend time…	positive
3	Basically there's a family where a little boy…	negative
4	Petter Mattel's "Love in the Time of Money" is…	positive
5	Probably my all-time favorite movie, a story…	positive
6	I sure would like to see a resurrection of a …	positive
7	This show was an amazing, fresh and innovative…	negative
8	Encouraged by the positive comments about this…	negative
9	If you like original gut wrenching laughter …	positive

FIGURE 4.7: A sample of some unprocessed reviews from the IMDb dataset.

The main advantage of having the data properly pre-processed: to reduce the dimensionality, enhance the accuracy of the proposed classifier and speed up the classification process, hence aiding in real-time sentiment analysis. Oppositely to the proposed system for sentiment analysis of tweets, in which two types of pre-processing were needed; Twitter specific pre-processing and standard pre-processing were performed, here only basic pre-processing was applied. In this step, words in each review were tokenised (separated) and punctuation was removed. Afterwards, we used the in-built Python *nltk* package for stemming, lemmatising and stopwords removal from each review. This phase consists on preparing the data for further processing following these steps:

1. Remove numeric and empty texts
2. Remove punctuation from texts
3. Convert words to lower case
4. *Remove stopwords:* as demonstrated in example 1 the dataset used obviously contain a lot of non relevant data (noise). Therefore, very basic and rudimentary cleanup needs to be performed. Arbitrary characters and other useless information such as punctuation, stop words, special characters and finally links/URL were removed since we found no significance in our classification approach.

   Example 1 "I was blessed to have seen this movie last night. It made me laugh, it made me cry and it made me love life. This movie is a great movie that depicts a love of a father for his son. Will Smith did an incredible job and deserves every accolade available to him. His son also did a fantastic job. There is a great lesson that is learned in this movie and it truly shares the struggles of everyday life. This movie was heart felt and touching. It was truly an experience worth having. Thank you for making this movie and I look forward to seeing it again."

   "blessed night made laugh made cry made love life great depicts love father son incredible job deserves accolade son fantastic job great lesson learned shares struggles everyday life heart felt touching experience worth making forward"

5. *Stemming and Lemmatisation:* are text normalization (or sometimes called word normalization) techniques. In our work as shown below in example 2 in

order to stem and lemmatise words and sentences we used the publicly available Python *nltk* package which is the Natural Language Toolkit package provided by Python for NLP tasks (Nahili et al., 2019)

Example 2 "Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured,[1][2] similar to data mining."

Stemmed sentence: "data scienc is an interdisciplinari field that use scientif method, process, algorithm and system to extract knowledg and insight from data in variou form, both structur and unstructur, [1][2] similar to data mine."

4.3.4 Feature selection

Feature selection is necessary when building a model in order to reduce overfitting, improve accuracy and reduce training time. We carry out our sentiment analysis proposal implementing several feature selection algorithms, one using the data representation method word2vec as a lexical resource (Chapter 2 subsection 2.8.3), the others building features using our training data with some additional variations. We use uni-grams, tf-idf, negation, stemming and stop words to extract useful features from the review corpus and build a feature vector for each review. Since we are training, testing and comparing the performance of our model against models from previous work on sentiment analysis of text. The used features were regarding words or sequence of words from these reviews. The use of uni-grams has been quite successful on reducing dimentiality in previous research (Wang et al., 2011, Mesnil et al., 2014, Tiwari et al., 2017), therefore, they were used as a feature in the proposed model. The advantage of uni-gram model is; each unique word in the pre-processed review corpus is considered as a feature, therefore, building a feature vector for a review is straightforward. First, a dictionary of all the words occurring in the review corpus is created. Then the frequency of occurrence of a word in a review is determined and stored in a word-review matrix. Finally, we apply the TF-IDF weighting technique to this matrix to obtain the final feature matrix.

Experimenting with an existing resource in the feature set offers a huge advantage in training time because there is no looping over the dataset. In addition, word2vec learns quickly compared to other models for it consists exclusively of two-layer neural networks that are trained to reconstruct linguistic contexts of words with no need for labels. The disadvantage resides on the fact that the features used are not extracted from the dataset, thus, there is a possibility of including insignificant features while significant features are not selected. For example, text reviews are known for containing words that are spelled wrong, but still hold sentiment information. As result of using such a small set of features it may cause the problem of high bias. The feature set was built using training data results in a larger feature set. It is more efficient as it selects only relevant features from the IMDb dataset itself, and therefore, improves the performance significantly. However, when manipulating large datasets such as IMDb; - going through the training set in a loop to select relevant features can be slow when the training size becomes large; - if we use a small training set though, the features selected are not representative of the entire dataset because they might have high bias;- the problem of high variance with a large feature set; in other words, while the training error reduces with a larger training set, the test error remains high.

4.4 Experimental setup

Our approach was implemented using the following environment (Table 4.1), we created a setup with the succeeding system requirements. We tested our proposed emb-CNN model on Windows platform. We used a Dell XPS 13 with Windows 10 core-i5 (64 bit) machine equipped with 8GB of RAM. The general requirements shown in Table 4.1 below. The tools and technology used as follow:

TABLE 4.1: General system requirements for our deep learning approach for sentence level sentiment analysis of IMDb movie reviews.

Component	**Type**
Operating system	Windows 10
Processor	Intelcore-i5 (64 bit)
Memory (RAM)	8GB
Software and Third-party tools	Spyder,NLTK,scikit-learn,Numpy,Matplotlib,TensorBoard

Python 3.6[1] (implementation language): Python is a general purpose, interpreted high level programming language whose design philosophy emphasises code readability. Its syntax is clear and expressive. Python has a large and comprehensive standard library and more than 25 thousand extension modules.

[1]https://www.python.org/

Spyder 3.6.4[2] **(development environment):** is a powerful scientific environment written in Python, for Python, and designed by and for scientists, engineers and data analysts. It offers a unique combination of the advanced editing, analysis, debugging, and profiling functionality of a comprehensive development tool with the data exploration, interactive execution, deep inspection, and beautiful visualization capabilities of a scientific package.

NLTK 3.4.5 [3] **(language processing modules):** The Natural Language Processing Toolkit is an open source language processing module of human language in python. Created in 2001 as a part of computational linguistics course in the Department of Computer and Information Science at the University of Pennslyvania. NLTK provides in-built support for easy-to-use interfaces over 50 lexicon corpora. NLTK was designed with four goals in mind:

1. *Simplicity:* Provide and intuitive framework along with substantial building blocks, giving users a practical knowledge of NLP without getting bogged down in the tedious house-keeping usually associated with processing annotated language data.

2. *Consistency:* Provide a uniform framework with consistent interfaces and data structures, and easily guessable method names.

3. *Extensibility:* Provide a structure into which new software modules can easily by accommodated, including alternative implementations and competing approaches on the same task.

4. *Modularity:* Provide components that can be used independently without needing to understand the rest of the toolkit.

Scikit-learn 0.22.1[4] **(building the model):** The scikit-learn library is an open source machine learning library that supports supervised and unsupervised learning. It also provides various tools for model fitting, data pre-processing, model selection and evaluation, and many other utilities.

NumPy v1.18[5] **(data manipulation):** It is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array

[2]https://www.spyder-ide.org/
[3]https://www.nltk.org/install.html
[4]https://scikit-learn.org/stable/
[5]https://numpy.org/doc/1.18/reference/index.html

object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical and much more. We use Numpy data structures because it performs better in terms of: Size - Numpy data structures take up less space; Performance - they have a need for speed and are faster than lists; Functionality- SciPy and NumPy have optimized functions such as linear algebra operations built in.

TensorBoard 2.1.0[6] (tracking experiment metrics): is a tool for providing the measurements and visualisations needed during the machine learning workflow. It enables tracking experiment metrics like loss and accuracy, visualizing the model graph, projecting embeddings to a lower dimensional space, and much more.

Matplolib 3.1.3[7] (plotting): It is a Python 2D plotting library which produces publication quality figures in a variety of hard copy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and four graphical user interface toolkits.

4.5 Experimental results and discussion

The proposed emb-CNN model is powerful in terms of feature extraction, which is very interesting in sentiment analysis, especially when dealing with long reviews where the extraction of its features remains difficult. The data was split evenly with 25,000 reviews intended for training and 25,000 for testing. Moreover, each set has 12,500 positive and 12,500 negative reviews. The model was configured as follows:

- The maximum length of vectors that can be created from a review was 300;
- The maximum review length in the convolutional neural network was 1600;
- We only consider the top 10,000 most commonly occurring words in the dataset;

1. *Embedding layer:* this is an important step in the proposed network architecture because it leads to initialise the parameters at a good point. It specifies the following arguments as input: top-words, embedding-vector-length and max-review-length.

[6]https://pypi.org/project/tensorboard/
[7]https://matplotlib.org/

2. The three convolution layers were as follows:

   - The first one applies a number of convolution kernels of size 64 with a filter-length of 3 and padding equal to "same"
   - The second one applies a number of convolution kernels of 32 with filter-length of 3 and padding equal to "same"
   - The third one applies a number of convolution kernels of 16 with filter-length of 3 and padding equal to "same"

   In order to preserve information and conserve the original size of the input sentence matrix, we used padding in every convolution layer.

3. *Flatten layer:* converts the output of the convolutional layers into a 1D vector, to be used by the fully activated layer (dense) for the final classification.

4. Regularisation was applied with a dropout of (0.2)

5. There was a fully connected layer.

6. The "sigmoid" activation function allowed linking the obtained results with the appropriate class.

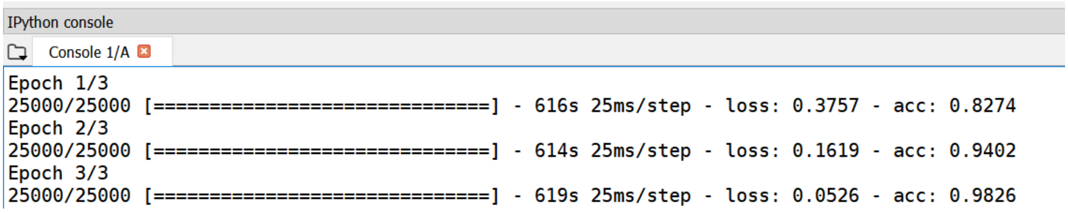
7. The loss function is indispensable for compiling the "binary-cross-entropy" model. We used the binary one as there are two classes in integer and categorical format (positive and negative).

8. The optimizer was "Adam".

9. We used accuracy as an evaluation metric (See subsection 2.8.4). Accuracy refers to the proportion of correct predictions made by the model.

$$Accuracy = \frac{Predictions}{TotalPredictions} \tag{4.1}$$

We benchmark the proposed emb-CNN model for sentiment analysis against prior models from the related work using the same IMDb dataset. Despite the large size of the dataset and the number of parameters in the network (Table 4.3) our model's strengths are its accuracy and training time which takes around fifteen to twenty minutes as shown in Figure 4.8.



```
IPython console
Console 1/A
Epoch 1/3
25000/25000 [==============================] - 616s 25ms/step - loss: 0.3757 - acc: 0.8274
Epoch 2/3
25000/25000 [==============================] - 614s 25ms/step - loss: 0.1619 - acc: 0.9402
Epoch 3/3
25000/25000 [==============================] - 619s 25ms/step - loss: 0.0526 - acc: 0.9826
```

FIGURE 4.8: The variation of loss and accuracy functions in each epoch during training.

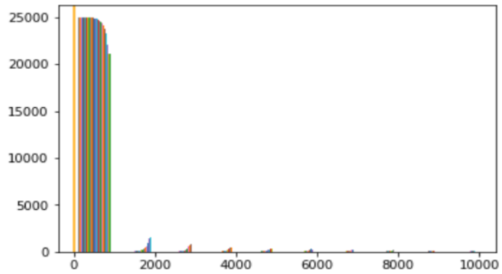


FIGURE 4.9: Training history of the model where 25k reviews were used for training and only top 10,000 most commonly occurring words in the dataset were used.

TABLE 4.2: The different configurations used to train and test our emb-CNN model.

Word embedding	Regulariser	Epoch	Batch-size	Optimizer	Accuracy
word2vec	Dropout	3	32	Adam	**85.99%**
word2vec	Dropout	3	64	Adam	85.80%
word2vec	Dropout	5	32	Adam	84.49%
word2vec	Dropout	5	64	Adam	85.98%
word2vec	Dropout	8	32	Adam	85.33%
word2vec	Dropout	8	64	Adam	85.46%

We compared the performance of our model under different configurations (see Table 4.2) for processing and analysing reviews text. Thus, the accuracy was computed according to three different iterations (i.e., 3, 5 and 8 epochs respectively), between two values of the batch-size (i.e., 32 and 64), and the optimizer Adam (Table 4.4, the highest value is highlighted in bold). The emb-CNN model gave the best performance under the configuration (3 epochs, batch size 32, and optimizer "Adam") that reached 85.99% in terms of accuracy (Table 4.4). Although convolutional neural networks extract high-level features in image analysis, our model performs well in 2D problems and trains in a short amount of time as shown in Figure 4.8.

TABLE 4.3: Total number of trainable parameters in the proposed emb-CNN model.

Layer type	Input shape	Parameter #
Embedding (word2vec)	(none, 1600, 300)	3000000
Convolution layer-1	(none, 1600, 64)	57664
Convolution layer-2	(none, 1600, 32)	6176
Convolution layer-3	(none, 1600, 16)	1552
Flatten layer	(none, 25600)	0
Dropout layer-1	(none, 25600)	0
Dense layer-1	(none, 180)	4608180
Dropout layer-2	(none, 180)	0
Dense layer-2	(none, 1)	181

With approximately 7 millions trainable parameters (Table 4.3) overfitting the training data is a serious problem, thus, dropout of (0.2) was applied. For reinforcing the generalisation power, we disabled the network with holes during training; this way the network is forced to build new paths and extract new patterns.

TABLE 4.4: Performance results of the proposed model compared to prior models in terms of accuracy.

Author/year	**Dataset**	**Highest accuracy**
Ouayang et al. (2015)	rottentomatoes dataset	45.4%
Tripathy et al. (2016)	*IMDb dataset*	70.16%-86.23%
Houshmand (2017)	SST dataset	40.5%-46.4%
Miedema (2018)	IMDb dataset	86.74%
Our emb-CNN model	IMDb dataset	**85.99%**

- **Figure 4.10 and 4.11** under the configuration: regulariser "dropout", epochs=3, batch-size=32, optimizer "Adam".
- **Figure 4.12 and 4.13** under the configuration: regulariser "dropout", epochs=3, batch-size=64, optimizer "Adam".
- **Figure 4.14 and 4.15** under the configuration: regulariser "dropout", epochs=5, batch-size=32, optimizer "Adam".
- **Figure 4.16 and 4.17** under the configuration: regulariser "dropout", epochs=5, batch-size=64, optimizer "Adam".
- **Figure 4.18 and 4.19** under the configuration: regulariser "dropout", epochs=8, batch-size=32, optimizer "Adam".

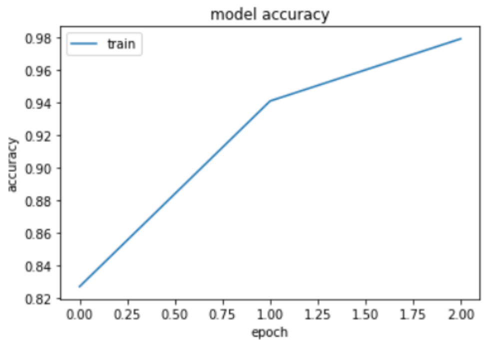


FIGURE 4.10: Accuracy plot of the emb-CNN model "3","32"

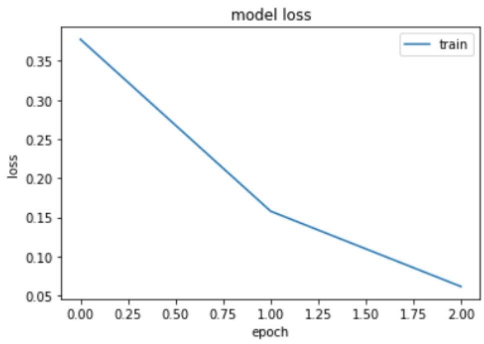


FIGURE 4.11: Loss plot of the emb-CNN model "3","32"

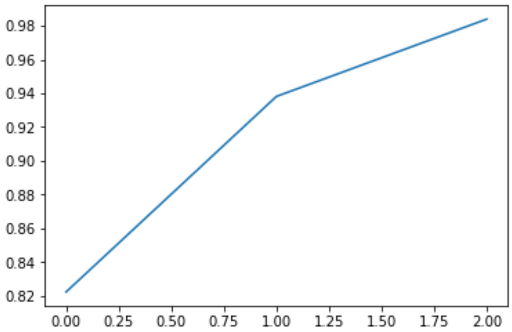


FIGURE 4.12: Accuracy plot of the emb-CNN model "3","64"

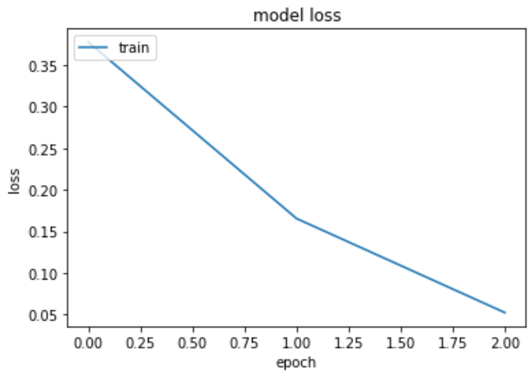


FIGURE 4.13: Loss plot of the emb-CNN model "3","64"

- **Figure 4.20 and 4.21** under the configuration: regulariser "dropout", epochs=8, batch-size=64, optimizer "Adam".

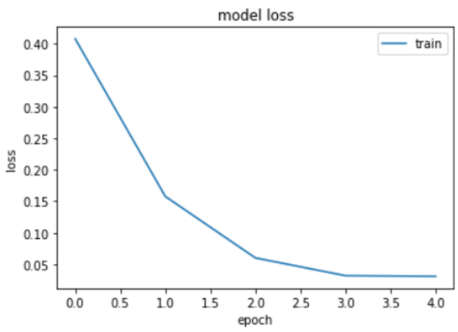


FIGURE 4.14: Accuracy plot of the emb-CNN model "5","32"

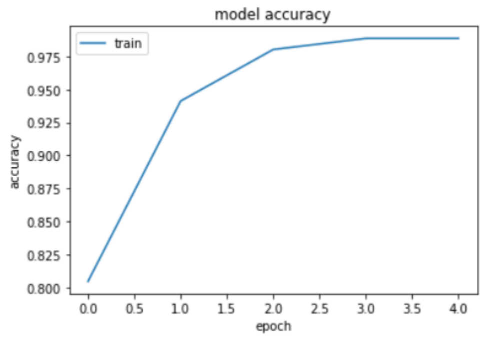


FIGURE 4.15: Loss plot of the emb-CNN model "5","32"

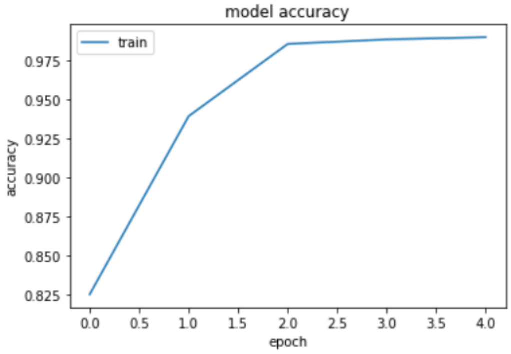


FIGURE 4.16: Accuracy plot of the emb-CNN model "5","64"

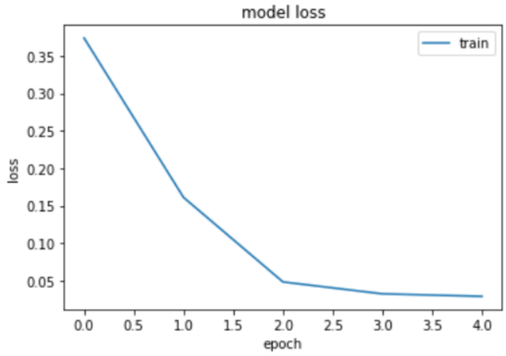


FIGURE 4.17: Loss plot of the emb-CNN model "5","64"

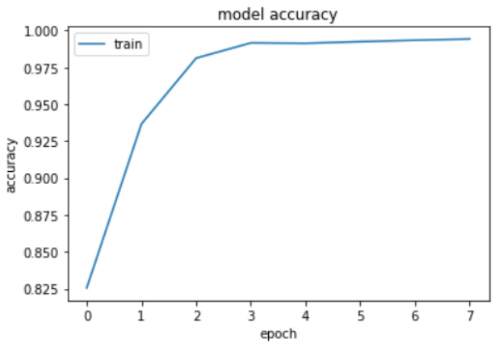


FIGURE 4.18: Accuracy plot of the emb-CNN model "8","32"

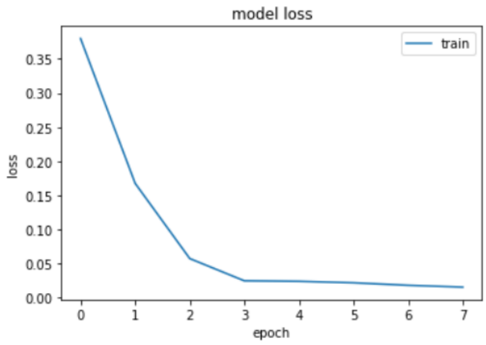


FIGURE 4.19: Loss plot of the emb-CNN model "8","32"

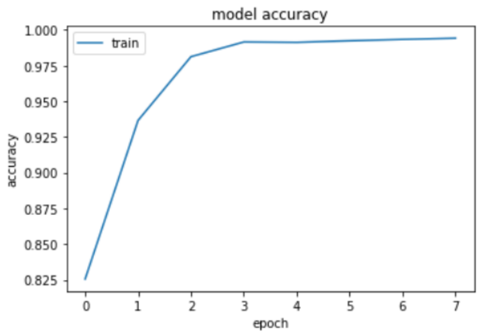


FIGURE 4.20: Accuracy plot of the emb-CNN model "8","64"

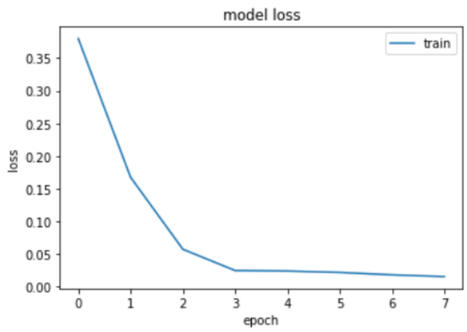


FIGURE 4.21: Loss plot of the emb-CNN model "8","64"

Despite the strong empirical performance achieved by the proposed network with 86.74% in (Miedema, 2018), and the good performance of the proposed classifiers (i.e. NB, ME, SGD, and SVM) in the work of (Tripathy et al., 2016), in addition to CNN networks experimented with in both (Ouyang et al., 2015, Houshmand, 2017), several conclusions were drown; First, based on the research of (Miedema, 2018), our proposed emb-CNN model is more robust where it has 7,673,753 parameters to be trained on, against only 373,301 in theirs. Also, the training time of our network is around 15 to 20 minutes versus 1 hour for their model, which proves the efficiency of our model. Other than this, there is no room for discussion nor comparison in terms of performance, due to the fact that the author implemented a recurrent neural network combined with LSTM architecture for sentiment analysis of IMDb reviews, and we propose a CNN model with an embedding layer architecture (word2vec). Second from the study of (Tripathy et al., 2016), they did not experiment with convolutional neural networks for sentiment classification, however we can point out that with an accuracy score of 85.99%, our emb-CNN model outperformed their SVM (70.16%), ME (83.36%) and SGD (83.36%) classifiers consecutively. Additionally, on one hand, in (Ouyang et al., 2015) their word2vec+CNN system for analysing online movie reviews had no room for improvement because the corpus that was used to conduct sentiment analysis was crawled from rottentomatoes movie reviews dataset which does not contain enough data to train a robust model. Also, their model obtained satisfactory results with 45.4% accuracy. In contrast, our proposed model was trained, tested and validated using a large-scale corpus the IMDb dataset with 50k reviews, besides with our emb-CNN network we were able to improve the accuracy by 40%. On the other hand, in comparison with the reported results in (Houshmand, 2017), the performance of their CNN model varied from 40.5% to 46.4% in terms of accuracy, while ours achieved 85.99%. Lastly, shuffling the training and testing sample sizes while experimenting with data representation methods like word embeddings, in our case word2vec significantly improved the network's performance up to 86.48%.

4.6 Conclusion

This chapter was divided into two main sections: the first one being our proposed deep learning approach by introducing a new convolutional neural network with an embedding layer using word2vec to predict the sentiment of movie reviews. The second one consisted of a detailed presentation of the network which was composed of several layers. We used word2vec as an embedding layer leading to initialise the parameters at a good point of the network. The model was trained using different features such as tokenisation, stemming and TF-IDF ..etc. Afterwards, the network was trained, tested and validated on the IMDb dataset using accuracy as an evaluation metric. We experimented with the proposed emb-CNN model under different configurations where binary-class sentiment analysis was performed. The proposed model has yielded better results as compared to previous networks. In summary, the proposed model for sentiment analysis of movie reviews has shown promising results during the evaluation phase. Consequently, it has provided supporting evidence for the accuracy and reliability of the proposed deep learning approach.

Chapter Five

Conclusion

5.1 Synopsis of the Thesis

Consisting of five chapters, this thesis presents a comprehensive and thorough research work in the field of sentiment analysis with main focus on the classification of social media text, which seeks to provide two approaches: the first one is a lexicon-based approach to conduct multi-class sentiment analysis by developing a sentiment lexicon, the second one is a deep learning based approach for binary-class sentiment analysis of IMDb movie reviews by proposing a novel convolutional neural network with an embedding layer. These sentiment analysis approaches have been proposed in order to answer the research questions and meet the objectives of this work. This thesis has undertaken the phases of design science methodology for information systems and software engineering (Wieringa, 2014), which include problem identification and objective definition as fundamental starting points (Chapter 3.3)

Chapter one has provided the research background and the growing demand for the development of sentiment analysis applications. Chapter two gave an in-depth literature review about prior research on sentiment analysis along with what challenges and problems were faced and why. By that, the problems encountered in the previous work and the research gaps have been identified, which also presented the motivation of this thesis. Lastly, research questions and goals have been defined. To recap, the research questions were as follows:

1. *How can online text data be automatically and accurately classified with respect to their sentiments?*
2. *How can the intensity of sentiment in online text data be effectively captured?*

3. *How can deep learning enhance the process of Data Analytics for companies so they can improve their decision making?*

In view of aiming to answer the above research questions, a sentiment analysis approach for fine-grained classification of tweets called TweetEcho has been proposed in chapter three followed by an enhanced CNN model for sentiment analysis of IMDb movie reviews. The sentiment classification task using TweetEcho is based on the sentiment lexicon that is manually constructed from the social media domain. The sentiment analysis task using the proposed emb-CNN model is carried out in a supervised manner i.e., the dataset used for analysis is a labelled one. The training data is used for training and based on this, the testing on testing data is carried out. The text polarity is collected and compared to the original label to obtain the accuracy.

Moreover, in order to prove that the proposed approaches work, the sentiment analysis system, called TweetEcho, has also been implemented based on the theoretical framework. The different phases of the proposed system were presented in chapter four. Chapter four has also demonstrated how TweetEcho performs sentiment analysis and the resulting outputs. Following that, the proposed emb-CNN model was built, trained and tested using IMDb dataset. In order to evaluate the model, we used accuracy as an evaluation metric. Comparison of the obtained results in this study with the reported performances in literature show that the model achieved better accuracy than previous convolutional neural network models.

The reported results in the phase of evaluation indicate that the proposed approach is able to automatically perform sentiment classification. The system is capable of detecting the intensity of sentiment and primarily relies on the manually developed sentiment lexicon, which has answered the first two research questions. Furthermore, the performance of the proposed deep learning approach has also been evaluated in chapter four. Each movie review from the IMDb dataset has been classified using the trained emb-CNN model, which has obtained an overall accuracy of 85.99% at sentence level analysis in this research, from which the third research question has also been answered.

5.2 Thesis contributions

In this section, the contributions that this thesis has made to the body of knowledge on sentiment analysis are highlighted, which encompass theoretical and methodological contributions. From a theoretical standpoint, the contribution of this research is the introduction and development of two approaches: (1) one for fine-grained Twitter sentiment analysis, and (2) a new CNN model for text analytics. These theoretical approaches have provided solutions to narrow the research gaps identified in prior studies (see Section 2.13). These approaches serve as a mechanism that can automatically classify social media text in order to provide directions for strategic decisions. Another theoretical contribution made by this research is that the lexicon-based approach has brought a dynamic process that follows the cycle of evaluation and refining (see Section 3.3.8), based on the design science methodology for information systems and software engineering (Wieringa, 2014). It allows researchers to refine the sentiment lexicon in order to improve the whole performance and take into consideration more terms (words), which is the limitation of other automatic generated sentiment lexicons.

As for the methodological contributions; the first methodological contribution is that the research has provided an effective way to separate the mix-sentiment text with text segmentation using the proposed lexicon-based approach, which is able to deal with real-world reviews accurately. Another methodological contribution that needs to be underlined is that the research has introduced an innovative way to process multi-word expressions by using the built sentiment lexicon, which contains not only adjectives, adverbs, but also includes nouns, verbs. Instead of treating text as a bag of words, the proposed approach is able to process multi-word expressions to gain more contextual information. A third methodological contribution of this research worth mentioning is the solution to current challenges faced in social media text analytics in regards to handling the linguistic phenomena such as negation, intensifiers, common misspellings and slang language by the addition multi-word terms to the sentiment lexicon. A fourth methodological contribution is this thesis has provided an efficient way to classify online text data using a deep learning approach by proposing a novel convolutional neural network model with an embedding layer. The model was built for analysing movie reviews and to do so, it was trained using a different selection of features under various configurations. The network achieved good results against the related work with 85.99% and, is able to identify the sentiment of real-world reviews accurately. Moreover, the short training time despite the large size of the dataset and the number of trainable parameters in the network. Furthermore, the source of data collection is also important. The dataset needs to be universally accessible which large number of researchers often consider for their analysis. In the present day scenario, Twitter data and movie reviews are an important source to analyse reviews. Thus, in this research we adopted both.

5.3 Affect of the results

The contributions which have been detailed in the previous section can have a large impact when applied in the real world. In the case of private and public enterprises, businesses can gain an insight into the sentiments (positive or negative) of customers and forecast their product or service by considering the sentiment analysis results. Twitter analysis provides businesses with both the sentiments of customers regarding a specific product and its strength, which offers these organisations detailed feedback about their own products and services or its competition. The positive insights can be used as inputs for market positioning against the competitors, product management and development, targeted marketing strategies to ensure that these product/service are not changed unnecessarily without great thought, which in return should increase the revenue of that product/service. The negative insights can be used as inputs to sales and strategy teams (e.g. pricing, what not to focus on in a competitive situation) and also for product management, research and development to understand the product or service, which their consumers were not satisfied with and see as requiring improvements. Although this thesis has examined the usage of sentiment analysis approaches in the domain of reviews (product and movie) and businesses providing products and services, there is no barrier to using the same approaches in other arenas, such as topic trending, collecting political sentiment and track political movements in countries for example the Hirak movement or the COVID19 pandemic in Algeria.

5.4 Limitations and Future Research

5.4.1 Limitations

The research comes as an attempt to find effective solutions to deal with the large volume of opinionated data automatically. First, a fine-grained sentiment analysis framework has not only been proposed, but also has been evaluated via the implemented system TweetEcho. Second, a deep learning approach was proposed based on a CNN model built to classify movie reviews, which was trained , tested and validated using the publicly available IMDb dataset. Although this thesis has answered the research questions , there are still some limitations in this work.

First of all, the development of the sentiment lexicon in this research requires human effort. Despite the fact that a manually built lexicon can achieve better accuracy, it is time-consuming. In this research, the number of sentiment terms in the current lexicon is limited, and keeping track of all new coming/trending terms on the social media domain is a challenging task because new words and expressions are created everyday in the speed of light. On the other hand, the different scores of semantic orientation that were manually assigned to the sentiment terms can be biased, which is also the common drawback of the lexicon-based approach for sentiment analysis.

A second limitation of this research worth mentioning is that the research mainly focuses on the informal text that is close to the way people speak in real life. Because the data that have been used in this research are tweets and movie reviews, which are user-generated content. The product reviews from other sources can be written in a formal way, such as professional critics or product reports from evaluative organisations are not taking into consideration in this work.

Another limitation would be regarding the fact that either approaches are implemented for conducting sentiment analysis on social media text, in which different reviews or comments contain symbols or images referred to as emoji's like **':)', ':('** which help in expressing the sentiment, but these images were removed during the pre-processing phase and thus, not taking into consideration for further analysis despite the fact that they convey sentiments. Additionally in order to give stress on a word, it is observed that some users often repeat some character in the word a number of times such as "greatttt, Fineee or amaaazing". These words usually do not have a proper meaning; but they may be considered and further processed to identify the intensity of sentiment associated with the given sentence to be analysed.

5.4.2 Suggestions for Future Research

The research in the field of sentiment analysis have been very active in recent years due to many challenging research problems and its practical applications. Although this research has bridged few gaps, further work is required for additional improvements. The most immediate enhancement for this research concerns the development of the sentiment lexicon, since the number of sentiment terms in the current lexicon is limited. Besides, the constructed sentiment lexicon is specific to online product reviews, thus it cannot be used in other domains such as movie domain. In addition, the techniques of generating lexicon automatically may be implemented in the future in order to conquer the limitation of this research, which can save time and human effort. Also, during this research it was noticed that the English language is mostly used in the sentiment analysis field. This research has only handled the English language. Properly, many data sources are built for this language. There is still lack of resources for the Arabic language. Even though the resources built for the Arabic language are not yet complete and not found easily as an open-source. On one hand, additional Algerian Arabic dialect sentiment lexicons can be developed, which can used in turn for further research for instance topic trending and collecting political sentiment towards the Hirak movement or the COVID19 pandemic in Algeria, so that the boundaries of the studies in the field of sentiment analysis can be driven forward. On the other hand, the proposed model can be adapted for sentiment analysis of Arabic text, include attention to the complexity of the Arabic language and its own characteristics to engineer useful features.

Another future work direction is to apply the proposed model architecture to other NLP applications such as sarcasm and spam filtering. As a sophisticated form of speech, sarcasm is always challenging in the field of sentiment analysis and NLP. Researches display that sarcastic text is not very common in product reviews, but more frequent in discussion related to other domains, such as politics (Liu, 2012). Although, in this research reviews text were analysed therefore, the data contain few sarcastic data, but still further research to explore this problem is still needed.

With the existence of a huge selection of supervised learning techniques in the fields of sentiment analysis and natural language processing, an additional future work suggestion is to explore other algorithms and conduct other experiments using Recursive Neural Network (RNN) with a Long Short-Term Memory (LSTM) architecture for sentiment categorisation of text review. Additionally, to investigate other word embeddings such as Glove (Pennington et al., 2014) and Fasttext (Bojanowski et al., 2017).

Lastly, in light of the discussed research future directions Cloud computing has been around for approximately two decades and Cloud adoption increases every year, since companies realise that it offers them access to world-class enterprise technology (Vedran, 2018). With aim of saving considerable amount of time for companies to download the necessary packages to run the application, we plan on implementing a Cloud solution and provide our application as a Software as a Service (SaaS) for enterprises. Since companies recognise Cloud computing benefits and are aware of its impact, they can utilise it to keep track of the impact of their products and services and hence, improve their production, revenues and stay ahead of their competitors.

Appendix One

Sentiment Lexicon

term	score	term	score	term	score
abandon	-2	amazing	2	admits	-1
admits	-1	attractive	2	anger	-3
admitted	-1	avoid	-1	angry	-3
adopt	1	awesome	4	angers	-3
adopts	1	awful	-4	annoy	-2
adorable	2	awkward	-2	annoys	-2
adore	2	bad	-3	annoyed	-2
adores	2	badly	-4	annoying	-2
adored	2	bamboozle	-2	applaud	2
advantage	2	bamboozled	-2	applauded	2
advantages	2	banish	-1	applauds	2
advantageous	2	banned	-1	applauding	2
afraid	-2	beautiful	4	appreciate	2
agonising	-3	beautifully	3	appreciated	2
agonising	-3	best	4	appreciating	2
agree	1	better	2	appreciation	2
amaze	2	big	1	approve	2
amazed	2	blah	-2	approves	2
amazes	2	blurry	-2	approval	2
abandoned	-2	shame	-2	abandons	-2
admire	2	astonished	3	abilities	1
admired	2	astounding	3	ability	1
admires	2	astoundingly	3	abuse	-2
admiring	2	attracted	2	abused	-2
admit	-1	attracting	2	abuses	-2
abusive	-2	accept	1	accepted	1
accepting	1	accepts	1	accident	-2
accidental	-2	accidentally	-2	accidents	-2

TABLE A.1: Terms in the sentiment lexicon for the social media domain.

term	score	term	score	term	score
accomplish	2	accomplished	2	accomplishes	2
admits	-1	attractive	2	admitted	-1
avoid	-1	adopt	1	awesome	4
adopts	1	don't love	-4	awkward	-2
adorable	2	stunned	4	don't hate	-3
advantage	2	very poor	-3	extremely poor	-4
advantages	2	banish	-1	banned	-1
advantageous	2	afraid	-2	OMG	1
agonising	-3	extremely beautiful	5	best	4
very big	2	agree	1	way better	3
amaze	2	big	1	much better	3
amazed	2	amazes	2	blurry	-2
bored	-2	boring	-2	bother	-2
bright	1	brightest	2	brightness	1
brilliant	4	broke	-1	broken	-1
calming	2	can't stand	-3	captivated	3
catastrophic	-4	clean	2	clearly	1
complain	-3	confuse	-2	confused	-2
confusing	-2	cool	1	cool stuff	1
crap	-3	crash	-2	damn	-3
damn it	-3	dead	-3	deception	-3
defect	-3	depressing	-2	desirable	2
desire	2	destroy	-3	destroyed	-3
destroying	-3	disappointed	-3	disappointing	-3
disappointment	-3	disaster	-5	dislike	-3
dissatisfied	-2	does not work	-3	dont like	-3
enjoy	2	enjoying	2	enjoyable	2
excellent	3	exiting	3	fail	-2
failing	-2	failure	-2	fascinated	3
fascinating	3	favorite	2	favourite	2
fine	2	flop	-2	fond	2
frustrated	-2	frustrating	-2	frustration	-2
fun	2	god	1	gud	1
goddamn	-3	good	3	great	3
greatest	3	ha	2	haha	3
hahaha	3	lol	3	happy	3
hate	-3	hated	-3	hoping	2
horrendous	-3	horrible	-3	horrific	-3
huge	1	inconvenient	-2	innovative	2
intelligent	2	interesting	2	joke	2
keen	1	lack	-2	lags	-2
lame	-2	lmao	4	lmfao	4
love	4	loved	4	low	-1
masterpiece	5	meaningless	-2	mess	-2

TABLE A.2: Terms in the sentiment lexicon for the social media domain.

term	score	term	score	term	score
nice	3	no fun	-2	no good	-4
not good	-4	not working	-2	perfect	5
perfectly	5	poor	-2	popular	3
positive	2	prblm	-2	regret	-2
remarkable	2	sad	-3	sadly	-3
satisfied	2	scam	-2	significant	2
stop	-1	stops	-1	success	2
successful	2	suck	-3	sucks	-3
terrible	-4	terribly	-4	unhappy	-3
upset	-2	want	1	waste	-1
weak	-2	win	3	wonderful	4
worse	-3	wow	4	usable	2
froze	-2	loose	-3	like	3
work	1	extremely nice	4	desperately	-3
outstandingly	5	outstanding	5	complaint	-3
happiness	3	fail	-2	damage	-3
afraid	-2	almost perfect	4	enjoys	2
amazing	4	barely usable	-3	break	-3
cannot install	-4	cracked	-3	damn good	5
don’t buy	-3	extremely	3	extremely good	5
fragile	-3	luv	4	no problem	1
not satisfied	-2	not perfect	-2	overheat	-4
overpriced	-4	worst	-4	perfect	5
really bad	-4	really love	5	really like	4
recommend	3	stunning	5	too bad	-4
trash	-3	ugly	-4	very bad	-5
best forgotten	-2	can’t be happier	3	not bad	2
very good	4	very much	3	loves	4
missing	-2	negative	-2	neglect	-2
affordable	1	afforded	1	affordability	1
goodness	3	sound	1	glad	3
pretty	2	nicely	3	oh	1
honestly	1	biggest	3	fast	2
faster	3	quick	2	quickly	2
quicker	3	loud	2	louder	3
loudest	3	dunno	1	nightmare	4
super	2	superb	3	poorly	-3
cheap	-1	cheapest	-2	cheaper	-1
flawless	4	ugh	1	ugliest	-4

TABLE A.3: Terms in the sentiment lexicon for the social media domain.

Appendix Two

Publications and Communications

Nahili W. and Dr. Rezeg K. *"Predicting Marketing Strategies with Social media: What can 140 characters reveal?"*, Poster presentation at LINFI Doctoral Day, 2017, Biskra, Algeria

Nahili W. and Dr. Rezeg K. *"A Cloud Computing approach for territorial intelligence based on social networks"*, Poster presentation at Journnées d'Etudes Informatique Théorique et Appliquées (JEITA), 2018, Biskra, Algeria

Nahili W., Dr. Rezeg K. and Miloudi L. *"Towards Better Decision-making with Twitter Sentiment Analysis"*, In proceedings of the 14th Edition of International conference on Business Intelligence and Big Data (EDA), 2018, Tangier, Morocco

Nahili W., Dr. Rezeg K. and Miloudi L. (2018) *"Towards Better Decision-making with Twitter Sentiment Analysis"*, in Journal of New Information Technologies, volume RNTI-B-14, pp. 31-40

Nahili W. and Dr. Rezeg K., *"Digital Marketing with Social Media: What Twitter Says!"*, in proceeding of IEEE of the 3rd International conference on Pattern Analysis and Intelligent Systems (PAIS), 2018, Tébessa, Algeria

Nahili W., Dr. Rezeg K. and Pr. Kazar O. *"Sentiment Analysis on Product Reviews Data Using Supervised Learning: A Comprehensive Review of Recent Techniques"*, in proceeding of ACM 8th International conference on Software Engineering and New Technologies, 2019, Hammamet, Tunisia

Nahili, W., Dr. Rezeg K. and Pr. Kazar O. (2019) *"A new corpus-based convolutional neural network for big data text analytics"*. Journal of Intelligence Studies in Business, 9(2), pp. 59-71, ISSN:2001-015X, DOI https://doi.org/10.37380/jisib.v9i2

Nahili W., Dr. Rezeg K. and Pr. Kazar O. (2020) *"Big Data Analytics using Supervised Learning: A Comprehensive Review of Recent Techniques"*, in International Journal for Research in Applied Science and Engineering Technology, 8(1), pp. 305-312, ISSN: 2321-9653

Bibliography

Statista. Information on the most popular networks worldwide. 2018. URL https://fr.statista.com/statistiques/570930/reseaux-sociaux-mondiaux-classes-par-nombre-d-utilisateurs/. Accessed 9 March 2020.

Emmanuel Pateyron. *La veille stratégique*. Paris, Economica, 1998.

Maud Pelissier and Isabelle Pybourdin. L'intelligence territoriale. entre structuration de réseau et dynamique de communication. *Les Cahiers du numérique*, 5(4):93–109, 2009.

Henri Matre. *Intelligence économique et stratégie des entreprises*. La documentation française, 1994.

D. K. Campbell, Bach D. R., A. Roepstorff, R. J. Dolan, and C. D. Frith. How the opinion of others affects our valuation of objects. *Current Biology*, 20(13):1165–1170, 2010.

K. Dave, S. Lawrence, and D Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *In Proceedings of World Wide web*, pages 519–528, 2003.

Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC06)*, May 2006. URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/384_pdf.pdf.

B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

Bing Liu. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139084789.

Tim Berners, Fischetti Mark, and Foreword Michael L. *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor.* Harper Information, 2000.

Tim O'Reilly. *O'reilly network: What is web 2.0.* 2005. URL http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-Web-20. Accessed 20 February 2020.

J. A. Obar and S. Wildman. Social media definition and the governance challenge: An introduction to the special issue. *Telecommunications Policy*, 39(9):745–750, 2015. ISSN 0308-5961. doi: https://doi.org/10.1016/j.telpol.2015.07.014. URL http://www.sciencedirect.com/science/article/pii/S0308596115001172.

S. George. *How Much Data Is Generated Every Minute On Social Media.* 2015. URL http://wersm.com/how-much-data-is-generated-every-minute-on-social-media/. Accessed 15 February 2020.

S. M. Mudambi and D. Schuff. What makes a helpful review? a study of customer reviews on amazon.com. *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, 34(1):185–200, 2010.

A. Y. Chong, B. Li, E. Ngai, E. Chng, and F. Lee. Predicting online product sales via online reviews, sentiments, and promotion strategies: A big data architecture and neural network approach. *International Journal of Operations and Production Management*, 36(4):358–383, 2016.

Q. Ye, R. Law, B. Gu, and W. Chen. The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior*, 27(2):634–639, 2011.

A. S. Cantallops and F. Salvi. New consumer behavior: A review of research on ewom and hotels. *International Journal of Hospitality Management*, 36:41–51, 2014.

K. Shrestha. *50 Stats You Need to Know About Online Reviews.* 2016. URL https://www.vendasta.com/blog/50-stats-you-need-to-know-about-online-reviews/. Accessed: 15 February 2020.

Zara Stone. *A Surprisingly Large Amount of Amazon Reviews Are Fake*. 2015. URL http://thehustle.co/. Accessed: 22 December 2019.

Graham Charlton. Ecommerce consumer reviews: why you need them and how to use them. 2012. URL Econsultancy.com.

Fan and Fuel. *No online customer reviews means BIG problems in 2017*. 2016. URL https://fanandfuel.com/no-online-customer-reviews-means-big-problems-2017/. Accessed 25 January 2020.

Luke Wallace. *Improving the Consumer E-commerce Experience Through Text Mining Industry View*. 2015. URL http://www.softwareadvice.com/bi/industryview/text-mining-report-2015/. Accessed 29 January 2020.

V. K. Jain and S. Kumar. Improving customer experience using sentiment analysis in e-commerce. *Handbook of Research*, 2016.

E. Alghamdi. The influence of social media on e-commence sites. 2013.

TripAdvisor. *TripAdvisor*. 2020. URL https://www.tripadvisor.fr/. Accessed: 15 January 2020.

Statista. *TripAdvisor Statistics and Facts*. 2019. URL https://www.statista.com/topics/3443/tripadvisor/. Accessed 15 February 2020.

S. Michael. *Social Media Marketing Industry Report*. 2012. URL http://www.socialmediaexaminer.com/SocialMediaMarketingIndustryReport.pdf. Accessed 29 October 2019.

P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *In Proceedings of the Association for Computational Linguistics*, pages 417–424, 2002.

Bing Liu. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, 2:627–666, March 2010.

Fuchun Peng, Dale Schuurmans, and Shaojun Wang. Language and task independent text categorization with simple language models. *Proceedings of the 2003 Conference of the North American Chapter of the Association for*

Computational Linguistics on Human Language Technology, 1:110–117, 2003. doi: 10.3115/1073445.1073470. URL https://doi.org/10.3115/1073445.1073470.

Xiaowen Ding, Bing Liu, and Philip S. Yu. A holistic lexicon-based approach to opinion mining. *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 231–240, 2008. doi: 10.1145/1341531.1341561. URL https://doi.org/10.1145/1341531.1341561.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37: 267–307, June 2011. doi: 10.1162/COLI_a_00049.

Saif Mohammad, S. Kiritchenko, and X. Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *CoRR*, abs/1308.6242, 2013. URL http://arxiv.org/abs/1308.6242.

Xi Ouyang, Pan Zhou, Cheng Li, and Lijun Liu. Sentiment analysis using convolutional neural network. *In IEEE International Conference on Computer and Information Technology*, pages 2359–2364, 2015. doi: 10.1109/CIT/IUCC/DASC/PICOM.2015.349.

D. G. Maynard and K. Bontcheva. Challenges of evaluating sentiment analysis tools on social media. *Proceedings of Language Resources and Evaluation Conference (LREC)*, pages 2359–2364, 2016. doi: 10.1109/CIT/IUCC/DASC/PICOM.2015.349.

S. M. Houshmand. Applications of deep learning to sentiment analysis of movie reviews. 2017. URL https://cs224d.stanford.edu/reports/Shirani-MehrH.pdf.

Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5:627–666, 2012. doi: 10.2200/S00416ED1V01Y201204HLT016.

G. Vinodhini and Dr Chandrasekaran. Sentiment analysis and opinion mining: A survey. *Int J Adv Res Comput Sci Technol*, 2, 2012.

Souneil Park. Sentiment classification using socio-linguistic clusters. *Workshop on Semantic Analysis TASS@ SEPLN*, pages 99–104, 2015a.

F. H. Khan, U. Qamar, and S. Bashir. Sentimi: Introducing point-wise mutual information with sentiwordnet to improve sentiment polarity detection. *Applied Soft Computing*, 39:140–153, 2016.

Python. *Python Language*. 2020. URL https://www.python.org/. Accessed: 15 January 2020.

Spyder. *Anaconda Environment*. 2020. URL https://www.spyder-ide.org/. Accessed: 15 January 2020.

NLTK. *NLTK: Natural Language processing ToolKit*. 2020. URL https://www.nltk.org/install.html. Accessed: 15 January 2020.

Numpy. *Numpy*. 2020. URL https://numpy.org/doc/1.18/reference/index.html. Accessed: 15 January 2020.

Tensorboard. *Tensorboard: TensorFlow's visualization toolkit*. 2020. URL https://pypi.org/project/tensorboard/. Accessed: 15 January 2020.

Matplotlib. *Matplotlib: Visualization with Python*. 2020. URL https://matplotlib.org/. Accessed: 15 January 2020.

IMDb movie reviews data set. URL https://www.imdb.com/.

Ajgt arabic jordanian twitter corpus. 2016. URL https://github.com/komari6/Arabic-twitter-corpus-AJGT. Accessed: 15 January 2020.

Andreas Kaplan and Michael Haenlein. Social media: back to the root sand back to the future. *Journal of Systems and Information Technology*, 14(2):101–104, 2012. ISSN 0308-5961. doi: DOI10.1108/13287261211232126.

Klaus Scherer. Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1):227–256, 2003.

Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. *In Proceedings of the 20th international conference on Computational Linguistics*, 39: 1367, 2004.

Michael Speriosu, Sudan Nikita, Upadhyay Sid, and Baldridge Jason. Determining the sentiment of opinions. *Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing*, 39:53–63, July 27-31 2011.

A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224n: Natural Language Processing with Deep Learning*, 1:12, 2009.

Nadia da Silva, R. Eduardo, Hruschka, and Hruschka Estevam. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, 66(1):170–179, 2014.

Hassan Saif, He Yulan, and Alani Harith. Semantic sentiment analysis of twitter. *The Semantic Web- ISWC 2012*, 66:508–524, 2012.

J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. *Proceedings of the Association for Computational Linguistics (ACL)*, 2007.

Akshat Bakliwal, Arora Piyush, Madhappan Senthil, Kapre Nikhil, Singh Mukesh, and Varma Vasudeva. Mining sentiments from tweets. *Proceedings of the WASSA*, 12, 2012.

Vishal Gupta and Gurpreet Lehal. A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, 12:60–76, 2009.

Jeff Zabin and Alex Jefferies. Social media monitoring and analysis: Generating consumer insights from online conversation. *Aberdeen Group Benchmark Report*, 2008. URL https://www.bibsonomy.org/bibtexkey/Zabin+Jefferies:08a/om.

S. Wei, W. Hongwei, and H. Shaoyi. Sentiment analysis of chinese microblogging based on sentiment ontology: a case study of 7.23 wenzhou train collision. *Connection Science*, 25(4):161–178, 2013.

D Sanjiv and C. Mike. Yahoo! for amazon: Extracting market sentiment from stock message boards. *In Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*, 2001.

N. Tetsuya and Y. Jeonghee. Sentiment analysis: Capturing favorability using natural language processing. *In Proceedings of the Conference on Knowledge Capture (K-CAP)*, 2003.

G. Anindya, G.I. Panagiotis, and S. Arun. Opinion mining using econometrics: A case study on reputation systems. *In Proceedings of the Association for Computational Linguistics (ACL)*, 2007.

Filho Balage and T. A. Pardo. Nilc usp: A hybrid system for sentiment analysis in twitter messages. *In Second Joint Conference on Lexical and Computational Semantics (SEM)*, 2:568–572, 2013.

M. Gamon, A. Aue, Oliver Corston, and E. Ringger. Pulse: Mining customer opinions from free text. *In Advances in Intelligent Data Analysis*, 5:121–132, 2005.

Janyce Wiebe, F. Rebecca, and O'Hara Thomas. Development and use of a gold-standard data set for subjectivity classifications. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 246–253, 1999. doi: 10.3115/1034678.1034721.

R. Narayanan, B. Liu, and A. Choudhary. Sentiment analysis of conditional sentences. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 1:180–189, 2009.

C. Long, J. Zhang, and X. Zhut. A review selection approach for accurate feature rating estimation. *In Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 1:766–774, 2010.

H. Wang, Y. Lu, and C. Zhai. Latent aspect rating analysis without aspect keyword supervision. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 618–626, 2011.

G. Li, R. Law, H. Vu, Rong Q., J., and X. R. Zhao. Identifying emerging hotel preferences using emerging pattern mining technique. *Tourism management*, 46: 311–321, 2015.

Emma Haddi. *Sentiment Analysis: Text pre-processing, reader views and cross domains*. Brunel Univesity- London College of Engineering-Department of Computer Science, 2015.

K. Anbananthen and A. Elyasir. Evolution of opinion mining. *Australian Journal of Basic and Applied Sciences*, 7(6):359–370, 2013.

A. Banfield. *Unspeakable Sentences: Narration and Representation in the Language of Fiction*. Boston: Routledge and Paul, 1982.

Marti Hearst. Direction-based text interpretation as an information access refinement. *Text-Based Intelligent Systems*, 1992.

J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210, 2005.

P. Shabajee and D. Reynolds. What is annotation? a short review of annotation and annotation systems. *ILRT Research Report*, (1053), 2012.

Claire Cardie. Empirical methods in information extraction. *Ai Magazine*, 18, 11 1997. doi: 10.1609/aimag.v18i4.1322.

C Cardie, J Wiebe, T Wilson, and D Litman. Combining low-level and summary representations of opinions for multi-perspective question answering. *Proceedings of the AAAI Spring Symposium on New Directions in Question Answering*, pages 20–27, 2003.

D. Jurafsky and J. H. Martin. *Speech and language processing (2nd edition) (prentice hall series in artificial intelligence) (2 ed.)*. Prentice Hall, 2008.

J. Yi, N. Tetsuya, B. Razvan, and N. Wayne. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2003.

B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. *Proceedings of the 14th International World Wide Web Conference*, pages 342–351, 2005.

S.M. Kim and E. Hovy. Automatic identification of pro and con reasons in online reviews. *Proceedings of the COLING/ACL Main Conference Poster Sessions*, pages 483–490, 2006. URL https://dl.acm.org/doi/10.5555/1273073.1273136.

A. Montejo-Ráez, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña-López. Ranked wordnet graph for sentiment polarity classification in twitter. *Computer Speech and Language*, 28(1):93–107, 2014.

B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *In Proceedings of Conference on Empirical Methods in NLP*, pages 79–86, 2002.

A Esuli and F. Sebastiani. Determining the semantic orientation of terms through gloss classification. *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, 2005.

P. Chaovalit and L. Zhou. Movie review mining: A comparison between supervised and unsupervised classification approaches. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS05)*, pages 112c–112c, 2005.

W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.

V. Hatzivassiloglou and K. McKeown. Predicting the semantic orientation of adjectives. *In Proceedings of 35th Meeting of the Association for Computational Linguistics*, pages 174–181, 1997.

J. Brooke. *A semantic approach to automated text sentiment analysis.* 2009.

R. Prabowo and M. Thelwall. Sentiment analysis: a combined approach. *Journal of Informetrics*, 3:143–157, 2009.

P. D. Turney and M. L Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

V. Hatzivassiloglou and J. M. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. *In Proceedings of the 18th conference on Computational linguistics*, 1:299–305, 2000.

P. J. Stone, D. C. Dunphy, and M. S. Smith. *The General Inquirer: a Computer Approach to Content Analysis.* Cambridge: The M.I.T. Press, 1966.

M. Hu and B. Liu. Mining and summarizing customer reviews. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1:168–177, 2004.

T. Luo, S. Chen, G. Xu, and J. Zhou. *Trust-based Collective View Prediction.* Springer, 2013.

Janyce Wiebe. Learning subjective adjectives from corpora. *In AAAI/IAAI*, pages 735–740, 2000.

Douglas Rice and Christopher Zorn. Corpus-based dictionaries for sentiment analysis of specialized vocabularies. *Political Science Research and Methods*, pages 1–16, 2013.

B. Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Computing Research Repository (CoRR)*, cs.CL/0409058:168–177, 2004. URL http://arxiv.org/abs/cs.CL/0409058.

Vikas Malik and Amit Kumar. Sentiment analysis of twitter data using naive bayes algorithm. *International Journal on Recent and Innovation Trends in Computing and Communication*, 6(4):120–125, 2018. ISSN 2321-169.

Bing Liu. *Opinion Mining and Summarization: Sentiment Analysis*. Tutorial given at WWW-2008, 2008. URL https://www.cs.uic.edu/~liub/FBS/opinion-mining-sentiment-analysis.pdf.

Jaap Kamps, Maarten Marx, Robert Mokken, and Maarten de. Using wordnet to measure semantic orientation of adjectives. *Proceedings of LREC*, 2004.

Saif Mohammad, Cody Dunne, and Bonnie Dorr. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 599–608, 2009. URL https://www.aclweb.org/anthology/D09-1063.

Soujanya Poria. Merging senticnet and wordnet-affect emotion lists for sentiment analysis. *IEEE International Conference on Signal Processing*, 2012. doi: 1251-1255.10.1109/ICoSP.2012.6491803.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *LREC*, 2010. URL http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf.

V. K. Jain and Y. Pandey. Analysis and implementation of sentiment classification using lexical pos markers. *International Journal of Computing, Communications and Networking*, 2(1):36–40, 2013.

Monica Malik, Habiba Sharib, and Agarwal Parul. A novel approach to web-based review analysis using opinion mining. *International Conference on Computational Intelligence and Data Science (ICCIDS 2018)*, 132:1202–1209, 2018.

L. Tan, J. Na, Y. Theng, and K. Chang. Sentence-level sentiment polarity classification using a linguistic approach. *Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation*, pages 77–87, 2011.

M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418, 2011.

P. Waila, V. K. Marisha, and M. Singh. Evaluating machine learning and unsupervised semantic orientation approaches for sentiment analysis of textual reviews. *IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore*, pages 1–6, 2012. doi: 10.1109/ICCIC.2012.6510235.

M. Ghiassi, Skinner J., and Zimbra D. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications*, 40(16):6266–6282, 2013. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2013.05.057.

J. Khairnar and M. Kinikar. Machine learning algorithms for opinion mining and sentiment classification. *International Journal of Scientific and Research Publications*, 3(6):1–6, 2013.

B. Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of the Association for Computational Linguistics (ACL)*, pages 115–124, 2005.

A. Aue and M. Gamon. Customizing sentiment classifiers to new domains: A case study. *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, 2005.

H. Yang, L. Si, and J. Callan. Knowledge transfer and opinion detection in the trec2006 blog track. *Proceedings of TREC*, 2006.

O. Kummer and J. Savoy. Feature selection in sentiment analysis. *Proceeding of the Conference en Recherche d'Infomations et Applications (CORIA)*, pages 273–284, 2012.

Agarwal Basant and Mittal Namita. *Prominent Feature Extraction for Sentiment Analysis*, volume 2. Springer International Publishing, 2016. doi: 10.1007/978-3-319-25343-5.

A. Abbasi, H. Chen, and A. Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions Information Systems*, 26(3):34, 2008. doi: 10.1145/1361684.1361685. URL http://doi.acm.org/10.1145/1361684.1361685.

V. Ng, S. Dasgupta, and S. M. Arifin. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. *Proceedings of the COLING/ACL on Main conference poster sessions*, 26(3):611–618, 2006.

D. Bespalov, B. Bai, Y. Qi, and A. Shokoufandeh. Sentiment classification based on supervised latent n-gram analysis. *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 375–382, 2011.

S. Srinivas, R. Sarvadevabhatla, K. Mopuri, and N. Prabhu. A taxonomy of deep convolutional neural nets for computer vision. *Frontiers in Robotics and AI*, 36(2), 2016.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Journal of Neural Computation*, 1(4):541–551, 1989.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *In proceeding of the IEEE*, 86(11):2278–2324, 1998.

P. Semanet, S. Chintala, and Y. LeCun. Convolutional neural networks applied to house numbers digit classification. *In Proceeding of the 21st International Conference on Pattern Recognition (ICPR)*, pages 3288–3291, 2012.

L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach. *Journal of Computer Vision and Image Understanding*, 106(1):59–70, 2007.

P. Semanet and Y. LeCun. Traffic sign recognition with multi-scale convolutional networks. *Proceeding of International Joint Conference on Neural Networks (IJCNN)*, pages 2809–2813, 2011.

A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *In Advances in neural information processing systems*, pages 1097–1105, 2012.

A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. *The International Conference on Language Resources and Evaluation (LREc)*, 10:1320–1326, 2010.

Y. Lin, J. Zhang, X. Wang, and A. Zhou. An information theoretic approach to sentiment polarity classification. *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, 10:35–40, 2012.

Yoon Kim. Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, 2014.

L. Dhande and G. Patnaik. Analyzing sentiment of movie review data using naive bayes neural classifier. *Int J Emerg Trends Technol Comput Sci*, 3:313–320, 2014.

L. Zhang, R. Ghosh, M. Dekhil, M. C. Hsu, and B. Liu. Combining lexicon-based and learning-based methods for twitter sentiment analysis. *HP Laboratories*, HPL-2011-89, 2011. URL http://www.hpl.hp.com/techreports/2011/HPL-2011-89.html.

Nadia Félix Felipe da Silva, F. S. Luiz, Coletta, R. Eduardo, Hruschka, R. Estevam, and Hruschka. Using unsupervised information to improve semi-supervised tweet sentiment classification. *Journal of Information Science*, pages 348–365, 2016.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the omg! *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

W. Nahili, K. Rezeg, and O. Kazar. Big data analytics using supervised learning: A comprehensive review of recent techniques. *International Journal for Research in Applied Science and Engineering Technology*, 8(1):305–312, 2020. ISSN 2321-9653. doi: 10.22214/ijraset.2020.1056.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. Sentiment analysis in twitter. *SemEval-2014 Task 9*, pages 73–80, 2014. doi: 10.3115/v1/S14-2009.

Pollyanna Goncalves, Matheus Araujo, FabrBozcio Benevenuto, and Meeyoung Cha. Comparing and combining sentiment analysis methods. *Proceedings of the 2013 Conference on Online Social Networks*, pages 27–38, 2013. doi: 10.1145/2512938.2512951.

A. Severyn and A. Moschitti. Twitter sentiment analysis with deep convolutional neural networks. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 959–962, 2015.

Paul Debjyoti, Li Feifei, Krishna Murali, Yu Xin, and Frost Richie. Compass: Spatio-temporal sentiment analysis of us election. *KDD 17*, pages 13–17, 2017.

Mubarok Mohamad, Adiwijaya, and Dwi Aldhi Muhammad. Aspect-based sentiment analysis to review products using naïve bayes. *AIP Conference Proceedings*, 2017. doi: https://doi.org/10.1063/1.4994463.

K. M. Alomari, H. M. ElSherif, and K. Shaalan. Arabic tweets sentimental analysis using machine learning. *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 602–610, 2017.

Ankit, Nabizath, and Saleena. An ensemble classification system for twitter sentiment analysis. *International Conference on Computational Intelligence and Data Science (ICCIDS 2018)*, 132:937–946, 2018.

S. Wang and C. D. Manning. Baselines and bigrams: Simple, good sentiment and topic classification. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers*, 2:90–94, 2012.

Dai Quoc Nguyen, Dat Quoc Nguyen, Thanh Vu, and Son Bao Pham. Sentiment classification on polarity reviews: an empirical study using rating-based features. *Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 128–135, 2014.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1:142–150, 2011. URL https://www.aclweb.org/anthology/P11-1015.

G. Mesnil, T. Mikolov, M. A. Ranzato, and Y. Bengio. Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews. *Workshop contribution at International Conference on Learning Representations ICLR*, 2014. URL arXivpreprintarXiv:1412.5335.

A. Tripathy, A. Agrawal, and S. K. Rath. Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, 57: 117–126, 2016.

P. Tiwari, B. K. Mishra, S. Kumar, and V. Kumar. Implementation of n-gram methodology for rotten tomatoes review dataset sentiment analysis. *International Journal of Knowledge Discovery in Bioinformatics (IJKDB)*, 7(1):30–41, 2017.

Fenna Miedema. *Sentiment Analysis with Long Short-Term Memory networks*. Research paper in Business analytics, 2018.

D. Law, R. Gruss, and A. S. Abrahams. Automated defect discovery for dishwasher appliances from online consumer reviews. *Expert Systems with Applications*, 67: 84–94, 2017.

S. Z. Haider. *An Ontology-Based Sentiment Analysis: A case study (Dissertation)*. Master Degree Project in Informatics, University of Skövde, 2012.

M. R. Yaakub, Y. Li, A. Algarni, and B. Peng. Integration of opinion into customer analysis model. *Proceedings of IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology*, pages 164–168, 2012.

Lina Zhou and Pimwadee Chaovalit. Ontology-supported polarity mining. *JASIST*, 59:98–110, 01 2008. doi: 10.1002/asi.20735.

Yejin Choi and Claire Cardie. Learning with compositional semantics as structural inference for subsentential sentiment analysis. *Proceedings of the conference on empirical methods in natural language processing*, pages 793–801, 2008.

C. S. Khoo, S. B. Johnkhan, and J. C. Na. Evaluation of a general-purpose sentiment lexicon on a product review corpus. *International Conference on Asian Digital Libraries*, pages 82–93, 2015.

Andrea Esuli. The user feedback on sentiwordnet. *Computing Research Repository (CoRR)*, abs/1306.1343, May 2013. URL arXivpreprintarXiv:1306.1343.

Monalisa Ghosh and Animesh Kar. Unsupervised linguistic approach for sentiment classification from online reviews using sentiwordnet 3.0. *Int J Eng Res Technol*, 2 (9), 2013.

A. Cernian, V. Sgarciu, and B. Martin. Sentiment analysis from product reviews using sentiwordnet as lexical resource. *International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages WE–15, 2015.

G. Qiu, B. Liu, J. Bu, and C. Chen. Expanding domain sentiment lexicon through double propagation. *International Joint Conferences on Artificial Intelligence*, 9: 1199–1204, 2009.

W. Nahili and K. Rezeg. Towards better decision-making with twitter sentiment analysis. *Journal of New Information Technologies*, RNTI-B-14:31–40, 2018.

W. Nahili, K. Rezeg, and O. Kazar. A new corpus-based convolutional neural network for big data text analytics. *Journal of Intelligence Studies in Business*, 9(2):50–71, 2019. ISSN 2001-015X. doi: https://doi.org/10.37380/jisib.v9i2.

Matthiew Russell. *Mining the social web*. O'Reilly Media, 2011.

Twitter-Dev. Twitter developer docs. 2020. URL https://developer.twitter.com/en/docs. Accessed: 15 January 2020.

Roel Wieringa. *Design Science Methodology for Information Systems and Software Engineering*. Springer Verlag Berlin and Heidelberg GmbH and Corporation, 2014.

Xiaowen Ding and Bing Liu. The utility of linguistic rules in opinion mining. *Proceedings of the 30th ACM SIGIR conference on Research and development in information retrieval*, pages 811–812, 2007. doi: 10.1145/1277741.1277921. URL https://dl.acm.org/doi/10.1145/1277741.1277921.

Raksha Sharma and Pushpak Bhattacharyya. Detecting domain dedicated polar words. *Proceedings of the International Joint Conference on Natural Language Processing*, 2013.

S. Park. Sentiment classification using socio-linguistic clusters. *TASS@ SEPLNl*, pages 99–104, September 2015b.

W. L. Hamilton, K. Clark, J. Leskovec, and D. Jurafsky. Inducing domain-specific sentiment lexicons from unlabeled corpora. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2016:595, 2016.

F. A. Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. 2011. URL arXivpreprintarXiv:1103.2903.

A. Montejo-Ráez, E. Martínez-Cámara, M. T. Martin-Valdivia, and L. A. Urena-Lopez. Random walk weighting over sentiwordnet for sentiment polarity detection on twitter. *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 3–10, 2012.

M. Guerini, L. Gatti, and M. Turchi. Sentiment analysis: How to derive prior polarities from sentiwordnet. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2016, 2013. URL arXivpreprintarXiv:1309.5843.

R. Xiao and H. Tao. A corpus-based sociolinguistic study of amplifiers in british english. *Sociolinguistic studies*, 1(2):241–273, 2007.

W. Chamlertwat, P. Bhattarakosol, T. Rungkasiri, and C. Haruechaiyasak. Discovering consumer insight from twitter via sentiment analysis. *Journal of Universal Computer Science*, 18(8):973–992, 2012.

Prem Melville, Gryc Wojciech, and D. Lawrence. Richard. Sentiment analysis of blogs by combining lexical knowledge with text classification. *Proceedings of the*

15th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1275–1284, 2009. doi: https://doi.org/10.1145/1557019.1557156.

L. Lin, J. Li, R. Zhang, W. Yu, and C. Sun. Opinion mining and sentiment analysis in social networks: a retweeting structure-aware approach. *IEEE/ACM 7th International Conference on Utility and Cloud Computing*, pages 890–895, 2014.

S. Ray. Common machine leaning algorithms. 2015. URL https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/. Accessed: 15 April 2019.

Tomas Mikolov, Ilya Sutskever, Kai Chen, S. Greg, and Corrado. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 2013.

Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. *Proceedings of ACL 2014*, 2014.

W. Yih, X. He, and C. Meek. Semantic parsing for single-relation question answering. *In ACL Proceeding*, 2014.

A. Gibson and J. Patterson. *Deep Learning. Chapter 1: A review on machine learning.* O'Reilly Media, Incoporation, 2017.

S. H. Lai, Lepetit V., Nishino K., and Sato Y. Computer vision – accv 2016. *13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24*, 10112, 2017. doi: 10.1007/978-3-319-54184-6,183-204.

ML Glossary. Activation functions. 2018. URL https://ml-cheatsheet.readthedocs.io/en/latest/activation_functions.html. Accessed: 15 December 2018.

Y.A. LeCun, L. Bottou, G.B. Orr, and K.R. Müller. Efficient backprop. in neural networks: Tricks of the trade. *13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24*, 10112:9–48, 2012.

L. Bottou. Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT'2010*, pages 177–186, 2010.

D.P. Kingma and J. Adam Ba. A method for stochastic optimization. *Proceedings of the 3rd International Conference for Learning Representations*, 2015. URL arXiv2014,arXiv:1412.6980.

Ravi Parikh and Matin Movassate. Sentiment analysis of user-generated twitter updates using various classification techniques. 2009. URL https://nlp.stanford.edu/courses/cs224n/2009/fp/19.pdf.

G. Petz, M. Karpowicz, H. Fürschuß, A. Auinger, V. Stříteský, and A. Holzinger. Opinion mining on the web 2.0 – characteristics of user generated content and their impacts. *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data. HCI-KDD: Lecture Notes in Computer Science*, 7947, 2013.

Chisholm Andrew and Hoffmann Markus. Text mining and visualization: Case studies using open-source tools. *Data Mining and Knowledge discovery series*, 2016.

J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M Martin. Learning subjective language. *Computational Linguistics*, 30(3):277–308, 2004.

G. Li and F. Liu. A clustering-based approach on sentiment analysis. *Intelligent Systems and Knowledge Engineering (ISKE)*, pages 331–337, 2010.

H. Wu, Luk R., Wong K., and Kwok K. Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems*, 26(3):26–32, 2008.

Jeffrey Pennington, Richard Socher, and D. Manning Christopher. Glove: Global vectors for word representation. *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL https://nlp.stanford.edu/projects/glove/.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. URL http://dx.doi.org/10.1162/tacl_a_00051.

Bozicevic Vedran. Cloud computing benefits: 7 key advantages for your business. 2018. URL https://www.globaldots.com/blog/cloud-computing-benefits. Accessed: 15 January 2020.

Koppel Moshe and Schler Jonathan. The importance of neutral examples for learning sentiment. *Computational Intelligence*, 22(2):100–109, 2006.

M. Taboada, C. Anthony, and K. Voll. Methods for creating semantic orientation dictionaries. *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06)*, pages 427–432, 2006.

J. Steinberger, T. Brychcın, and M. Konkol. Aspect-level sentiment analysis in czech. *ACL*, 2014.

Chetan Mate. Product aspect ranking using sentiment analysis: A survey. pages 126–127, 2016.

A. S. Abrahams, Jia Jiao, G. A. Wang, and W. Fan. Vehicle defect discovery from social media. *Decision Support Systems*, 54:87–97, 2012. doi: https://doi.org/10.1016/j.dss.2012.04.005.

Kumar, Jain, and Pandey. Analysis and implementation of sentiment classification using lexical pos markers. *International Journal of Computing, Communications and Networking*, 2(1):36–40, 2013.

Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.

Vivek Yadav. How neural networks learn nonlinear functions and classify linearly non-separable data? *Medium*, 2017. URL https://medium.com/@vivek.yadav/. 19 January 2020.

Read Jonathon. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. *Proceedings of the ACL Student Research Workshop*, pages 43–48, 2009.

D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. Learning sentiment-specific word embedding for twitter sentiment classification. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 1:1555–1565, 2014. URL https://www.aclweb.org/anthology/P14-1146.pdf.

T. P. Szynalski. *Formal and informal English*. 2010. URL http://www.antimoon.com/how/formal-informal-english.html. 28 January 2020.

V. Niven. *Eliminate the noise from your social media streams*. 2012. URL http://www.needtagger.com/eliminate-the-noise-from-your-social-media-streams/. 13 January 2020.

G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Leonardo Marchi and Laura Mitchell. *Hands-On Neural Networks: Learn how to build and train your first neural network model using Python*. Packt Publishing, 2019.

Antonio Gulli and Sujit Pal. *Deep Learning with Keras: Implementing deep learning models and neural networks with the power of Python*. Packt Publishing, 2017.

Liu Bing, Messina Enza, Fersini Elisabetta, and Alberto Pozzi Federico. *Sentiment Analysis in Social Networks*. Morgan Kaufmann, 2016. ISBN 9780128044384. URL https://www.oreilly.com/library/view/sentiment-analysis-in/9780128044384/.

Indurkhya Nitin and Damerau Fred J. *Handbook of Natural Language Processing*. Chapman and Hall/CRC Machine Learning and Pattern Recognition. Chapman and Hall/CRC, 2010.

Liu Bing. *Web Data Mining*. Springer, 2011.

Sarkar Dipanjan. *Text Analytics with Python: A Practitioner's Guide to Natural Language Processing*. Apress, 2019.

B. Bengfort, R. Bilbro, and T. Ojeda. *Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning*. O'Reilly Media Inc USA, 2018.

Aurelien Geron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media Inc USA, 2017a.

Steven Bird, Ewan Klein, and Edward Lopez. *Natural Language Processing with Python*. O'Reilly Media, 2009.

Aurelien Geron. *Machine Learning avec Scikit-Learn: Mise en oeuvre et cas concrets*. Dunod, 2019.

Aurelien Geron. *Deep Learning avec TensorFlow: Mise en oeuvre et cas concrets*. Dunod, 2017b.

Wes McKinney. *Python for Data Analysis: Data Wrangling with Pandas, Numpy, and IPython*. O'Reilley Media Inc USA, 2017.

Jake Vanderplas. *Python Data Science Handbook*. O'Reilley Media, 2016.

Andreas Mueller and Sarah Guido. *Machine learning avec Python*. First Interactive, 2018.

M. Duwairi and Qarqaz Islam. Sentiment analysis in arabic tweets. *IEEE conference on Information and Communication Systems (ICICS)*, 2014.

Ait Hammou Badr, Ait Lahcen Ayoub, and Mouline Salma. Towards a real-time processing framework based on improved distributed recurrent neural network variants with fasttext for social big data analytics. *Information Processing and Management*, 57(1), 2020. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2019.102122.

Wilson Theresa, Wiebe Janyce, and Hoffmann Paul. Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354, 2005.

Salloum Said, Al-Emran A. Q., and Shaalan. Khaled. A survey of arabic text mining. *Intelligent Natural Language Processing: Trends and Applications*, pages 417–431, 2018.

Ryan Mitchell. *Web Scraping With Python: Collecting More Data from the Modern Web*. O'Reilly Media Inc USA, 2018.

Alice Zheng. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media Inc USA, 2018.

A. Basant and N. Mittal. Optimal feature selection for sentiment analysis. *Computational Linguistics and Intelligent Text Processing*, 7817:13–24, 2013. doi: https://doi.org/10.1007/978-3-642-37256-8_2.

Nicholls Chris and Song Fei. Comparison of feature selection methods for sentiment analysis. *Advances in Artificial Intelligence*, 6085:286–289, 2010. doi: https://doi.org/10.1007/978-3-642-13059-5_30.

S. KÜBLER, C. LIU, and Z. SAYYED. To use or not to use: Feature selection for sentiment analysis of highly imbalanced data. *Natural Language Engineering*, 24 (1):3–37, 2017. doi: 10.1017/S1351324917000298.

Duric Adnan and Song Fei. Feature selection for sentiment analysis based on content and syntax models. *Decision Support Systems*, 54(4):704–711, 2012. ISSN 0167-9236. doi: https://doi.org/10.1016/j.dss.2012.05.023.

P. H. Shahana and B. Omman. Evaluation of features on sentimental analysis. *Procedia Computer Science*, 64:1585–1592, 2015. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2015.02.088. Proceedings of the International Conference on Information and Communication Technologie.

D. Khanaferov, C. Luc, and T. Wang. Social network data mining using natural language processing and density based clustering. *IEEE International Conference on Semantic Computing(ICSC)*, pages 250–151, 2014.

D. A. Ostrowski. Semantic social network analysis for trend identification. *IEEE Sixth International Conference on Semantic Computing*, pages 215–222, 2012.

Ellen Riloff, Janyce Wiebe, and William Phillips. Exploiting subjectivity classification to improve information extraction. *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05)*, pages 1106–1111, 2005.

David Vilares, Alonso Miguel, and Gomez-Rodriguez Carlos. A syntactic approach for opinion mining on spanish reviews. *Natural Language Engineering*, 21(1): 139–163, 2015.

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

Salloum Said, Al-Emran Mostafa, and Shaalan. Khaled. A survey of text mining in social media: Facebook and twitter perspectives. *Advances in Science, Technology and Engineering Systems Journal*, 2(1):127–1133, 2017.

Bo Wang and Min Liu. Deep learning for aspect based sentiment analysis. *Commonications of the ACM*, 56(4):82–89, 2013. doi: doi:10.1145/2436256.2436274.

R. Feldman. Techniques and applications for sentiment analysis. 2015. URL https://cs224d.stanford.edu/reports/WangBo.pdf.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan Routledge, and Noah Smith. From tweets to polls: Linking text sentiment to public opinion time series. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 11:122–129, 2010.

Tang Duyu, Qin Bing, and Liu Ting. Deep learning for sentiment analysis: successful approaches and future challenges. *WIREs Data Mining Knowledge Discovery*, 5: 292–303, 2015. doi: 10.1002/widm.1171.

Erik Cambria, Schuller Bjrn, and Xia Yunqing. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21, 2013.

R. Agrawal and M. Batra. A detailed study on text mining techniques. *International Journal of Soft Computing and Engineering (IJSCE)*, 2(6):118–121, 2013. ISSN 2231-2307.

S. Poria, E. Cambria, D. Hazarika, and P. Vij. A deeper look into sarcastic tweets using deep convolutional neural networks. *In COLING*, 2016. URL https://arxiv.org/abs/1610.08815.

Amy Gesenhues. *Survey: 90% Of Customers Say Buying Decisions Are Influenced By Online Reviews*. 2013. URL https://marketingland.com/. Accessed 9 March 2020.

D. Tang and M. Zhang. Deep learning in sentiment analysis. *Deep Learning in Natural Language Processing*, pages 219–253, 2018.