MULTI-TIER CLASSIFICATION BASED ON SENTIMENT, TYPE, EMOTION AND PURPOSE FOR ONLINE DIABETES COMMUNITY

WANDEEP KAUR A/P RATAN SINGH

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2020

MULTI-TIER CLASSIFICATION BASED ON SENTIMENT, TYPE, EMOTION AND PURPOSE FOR ONLINE DIABETES COMMUNITY

WANDEEP KAUR A/P RATAN SINGH

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2020

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Wandeep Kaur A/P Ratan Singh

Matric No: WHA150005

Name of Degree: Doctor of Philosophy

Title of Thesis ("this Work"):

Multi-Tier Classification Based on Sentiment, Type, Emotion and Purpose For Online Diabetes Community

Field of Study: Opinion Mining

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date: 4th April 2020

Subscribed and solemnly declared before,

Witness's Signature Date: 11th April 2020

Name:

Designation:

MULTI-TIER CLASSIFICATION BASED ON SENTIMENT, TYPE, EMOTION AND PURPOSE FOR ONLINE DIABETES COMMUNITY

ABSTRACT

The evolution of social media platforms has created a niche for users to increasingly turn to such sites in order to share and exchange health related information. Facebook being one of the largest social networking sites has only encouraged such exchange thus mounting to a sheer amount of data that is hidden within unstructured text. The aim of this research is to propose a multi-tier classification based on sentiment, type, emotion and purpose (STEP) to classify data collected from diabetes community within Facebook. There are three tiers within the proposed STEP framework namely type, purpose and sentiment (and emotion within same tier). The first tier looks into the classification of type of diabetes. Here a manual type lexicon dictionary catering for all three forms of diabetes (type1, type 2 and gestational diabetes) was created. Naïve Bayes using n-gram was used for classification purpose where the proposed STEP framework was able to produce a F1-Score of 77% against benchmark models. Posts that could not be classified into any one type were grouped under Other while the correctly classified posts from this tier moved down to the next tier for purpose classification. In the next tier, posts were classified according to symptoms, lifestyle and treatment. A weighted information gain feature selection technique was adopted where weights were redistributed for those features that have been wrongly classified within the training phase. Co-training multinomial Naïve Bayes was used where the two base classifiers were used for both label and feature classification. The uniqueness lies in using dimensionality reduction technique of converting numeric vectors to string vectors using Word2Vec that improved F1-Score of 61% compared to only 48%. The last tier in the proposed STEP framework looked into sentiment and emotion classification. Here a mathematical equation was proposed to calculate sentiment intensity using Facebook behaviors of like, comment, share and reaction. Studies in the past have looked to analyze the use of this behaviors and how they impact sales, however, the attempt made in this research is to convert those numbers to intensity which could be used to better classify sentiment. Results show proposed sentiment classifier was able to produce better classification of F1-Score 84%. Emotion classification was also conducted within the same tier where Word2Vec common bag of words model was adopted using bootstrapping methodology. A similarity check between annotated corpus and Emolex determined the dominant emotion and thus classified post accordingly. This improved the classification process from detecting multiple emotion per post to classifying the most dominant emotion extracted from post. The proposed framework was able to improve overall classification accuracy within each of its tiers and using a multi-tier framework, it was able to remove posts that do not contribute towards classification within the upper layers thus contributing to a more refined dataset for classification within its lower tiers.

Keywords: multi-tier, sentiment, emotion, purpose, Facebook

RANGKA KLASIFIKASI BERBILANG PERINGKAT SENTIMENT, TYPE, EMOTION, DAN PURPOSE UNTUK KOMUNITI KENCING MANIS ATAS TALIAN

ABSTRAK

Evolusi platform social media telah memberi peluang kepada pengunna untuk menggunakan platform ini bagi tujuan berkongsi dan bertukar maklumat berkaitan kesihatan. Facebook sebagai rangkaian sosial terbesar telah menggalakan lagi pertukaran informasi yang menjana data lumayan yang tersembunyi di antara teks yang tidak berstruktur. Tujuan kajian ini adalah untuk megesyorkan rangka klasifikasi berbilang peringkat sentiment, type, emotion dan purpose (STEP) untuk mengklasifikasikan data yang dikumpul daripada komuniti kencing manis Facebook. Terdapat tiga peringkat dalam rangka klasifikasi STEP yang disyorkan iaitu jenis (type), tujuan (purpose) dan sentiment serta emosi yang terletak pada peringkat yang sama. Di peringkat pertama, sebuah lexicon manual disediakan khas untuk mengklasifikasi jenis kencing manis. Lexicon ini bertujuan untuk megklasifikasikan ketiga-tiga jenis kencing manis iaitu jenis 1, jenis 2 dan kencing manis semasa mengandung. Naïve Bayes menggunakan n-gram telah diadaptasi untuk tujuan pengelasan di peringkat ini. Hasil F1-Score 77% diperoleh terhadap model penanda aras. Pos Facebook yang tidak dapat diklasifikasikan dalam mana-mana jenis telah diklasifikasikan sebagai Lain-lain. Pos-pos yang berjaya diklasifikasi dengan betul dapat meneruskan perjalanan ke peringkat kedua dalam rangka klasifikasi STEP yang disyorkan iaitu rangka tujuan. Di rangka ini pos-pos telah diklasifikasikan dalam tiga kelas (tujuan) iaitu simptom, cara hidup dan rawatan. Dalam peringkat ini teknik information gain yang mempunyai pemberat digunakan di mana bagi pos-pos yang tidak dapat diklasifikasikan dengan betul, pemberat akan mengira semula nilai pemberat dan mengagihkan semula nilai tersebut bagi tujuan latihan. Algoritma latihan bersama (co-training) digunakan di mana dua klasifikasi asas untuk label dan ciri dimanipulasi. Uniknya ialah menukarkan vektor numerik kepada vektor tali untuk mengecilkan dimensi. Menggunakan kaedah ini, hasil F1-Score yang diperoleh ialah 61% berbanding 48%. Peringkat terakhir rangka klasifikasi STEP ialah mengklasifikasikan sentimen dan emosi. Bagi klasifikasi sentiment, kajian ini telah menggumpul nilai-nilai like, comment, share dan reaction yang terdapat pada Facebook dan mencadangkan formula matematik untuk menukar nombor-nombor tersebut kepada intensity sentimen untuk memperoleh klasifikasi yang lebih sempurna. Penggunaan nombor-nombor like, comment, share dan reaction telah membantu mencapai F1-Score yang lebih memberangsangkan. Akhir sekali dalam klasifikasi emosi, Word2Vec dan common bag of words model telah digunakan. bersama kaedah bootstrapping. Menggunakan kaedah ini, nilai persaamaan antara data anotasi dan emolex dikira dan emosi yang paling dominan akan diklasifikasikan berbanding dengan hanya menggunakam Emolex sahaja. Melalui kajian ini, didapati penggunaan klasifikasi berperingkat membantu menyingkirkan data-data yang tidak menyumbang terhadap proses pengelasan di peringkat atas dan menyediakan korpus yang mantap untuk pengelasan di peringkat bawah.

Keywords: berbilang peringkat, sentiment, emosi, tujuan, Facebook

ACKNOWLEDGEMENTS

To God

To my supervisor, Dr. Vimala Balakrishnan – you are everything and more. Your patience, your guidance, your trust, your time, your words of encouragement. Words may fail but the gratitude I have in my heart for you will go with me to my grave.

To my husband – this journey has been a testament to our marriage time and time again. Yet if I have reached this far, it's because you had a very huge role to play.

To my mother – in the toughest of times when I felt completely alone, I will forever be thankful I had you in my corner.

Last but not least, to my little one who left way too soon – there is not a day that goes by where I do not think of you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

AUC	:	Area Under Curve
CBoW	:	Common Bag of Words
DC	:	Dependency Classifier
EMOLEX	:	NRC Emotion Lexicon
FC	:	Feature Classifier
IG	:	Information Gain
LDA	:	Latent Dirichlet Allocation Model
LVS	:	Label Value Set
MNB	:	Multinomial Naïve Bayes
NB	:	Naïve Bayes
NLP	:	Natural Language Processing
POS	:	Part of Speech
SVM	:	Support Vector Machine
STEP	:	Sentiment-Type-Emotion-Purpose
SWN	:	SentiWordNet
WFS	:	Word Feature Set

CHAPTER 1: INTRODUCTION

Sentiment analysis or opinion mining is an arithmetic study focusing on extracting opinions, sentiments, evaluations, attitudes, moods and emotions from written text where it has been identified as the most active research area within the natural language processing, data mining as well as information retrieval domain (Cambria, Das, Bandyopadhyay, & Feraco, 2017).

Cambria et al. (2017) defines sentiment as an attitude, thought or judgement prompted by a feeling. By this definition, sentiment is nothing but an opinion enclosed with an emotion. Hence, when it comes to conducting a sentiment analysis study, it would also include exploring the emotion of the written text. It would be easy to muddle between sentiment and emotion as the two share a very strong correlation. For example, there are instances where emotion motivates an individual to deduce an entity and construct an opinion about it (Giatsoglou et al., 2017; M. S. I. Malik & Hussain, 2017). Similarly, an opinion of an individual can also stir emotions in another (Desmet & Hoste, 2013). Additionally, a written text can also imply contradicting opinions and emotions (Yadollahi, Shahraki, & Zaiane, 2017). For example, "*Although I hate avocadoes, I know it is good for my health*" portrays a negative emotion and positive opinion on the same entity, that is, the avocadoes. With respect to the above, Yadollahi et al. (2017) categorized sentiment analysis into two parts, namely opinion mining (expression of opinions) and emotion mining (articulation of emotion). Opinion mining is interested in classifying opinions as either positive, negative or neutral while emotion mining is concerned with detecting emotion within written text as joy, sadness, anger, fear, disgust etc.

Andreu-Perez, Poon, Merrifield, Wong, and Yang (2015) labeled social media platforms as core elements of social health as more users are referring to social media sites to seek information, share personal experiences regarding diseases, medical treatments as well as communicating with other patients who are going through similar experiences. Health organizations are aware on the existence and contribution of such online communities; however, it is a battle to extract useful observations from the huge volume of data (Abedin et al., 2017; McRoy et al., 2018). Therefore, it is essential to process the available unstructured data to map out relevant information that would eventually benefit the health sector, including patients, caregivers, doctors and pharmaceutical companies as well as government entities responsible for drawing up health related policies.

The rest of this chapter is organized as follows: a general idea of opinion and opinion mining is presented in the first section followed by a brief introduction into emotion analysis. Next, a look into sentiment analysis is presented before moving on to introduce this research which includes the problem statement, aim of research, objectives, research questions, research contribution and significance. Lastly an outline of the methodology adopted as well as overall thesis layout is also presented.

1.1 Opinion Mining

Opinion mining is referred to as a subdiscipline study that combines information retrieval and computer linguistics, but with a bigger focus on the opinions expressed within the document instead of the topic of the document (B. Liu, 2012). In this section, the definition of opinion and introduction to the components of opinion are discussed.

1.1.1 Opinion

Literature has defined opinion as a quadruple (g, s, h, t) where g represents the sentiment target, s is the sentiment of the opinion about the target, h is the opinion holder and t is the time the opinion was expressed (Cambria et al., 2017; B. Liu, 2012). The time factor (t) is significant as an opinion expressed a year ago may not be the same as the opinion expressed today. Similarly, the opinion holder (h) is an influential element as the opinion of an important stakeholder (e.g. Prime Minister) weighs more than someone on the street. The opinion target (g) on the other hand, is important for two reasons: firstly, it is crucial to identify the exact target for a positive or negative sentiment in the case of multiple targets. For example, "*The food is great but the ambiance is nothing to shout about*", suggests a positive sentiment for the food served but a negative sentiment for the ambiance of the restaurant. Secondly, phrases such as great, wonderful, terrible, amazing etc. used to express sentiments and opinion targets are known to have syntactic relations (Hu & Liu, 2004; Qiu, Liu, Bu, & Chen, 2011) which permits classification algorithms to extract sentiment terms as well as opinion targets; two vital elements needed for conducting sentiment analysis (Cambria et al., 2017).

The sentiment target (g) is an attribute or entity upon which the sentiment is expressed. For example, "*The display quality of a Samsung TV is crisp*", without knowing the display quality being referred to belongs to Samsung TV, the opinion of the sentence is pointless. B. Liu (2007) discovered an entity can be deconstructed and illustrated hierarchically where the said entity (e) represents a product, service, topic, person, issue or event. The author formulated it as e: (T, W), where T refers to the hierarchy of parts and W is a set of attributes of (e). For instance, (e) can refer to a model of a phone, e.g., iPhone 7 which has attributes such as weight, screen resolution, operating system, processor etc. The processor of iPhone7 can have its own attributes such as speed, RAM, etc.

1.1.2 Basic Components of Opinion

There are three basic components of an opinion, namely, opinion holder, object and opinion (Figure 1.1). These components are described as:

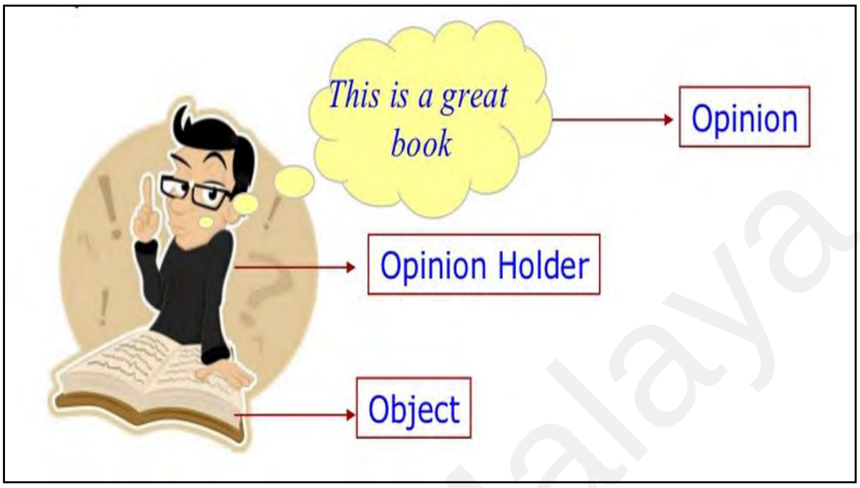


Figure 1.1 Components of an Opinion (Seerat & Azam, 2012)

1.1.2.1 Opinion Holder

The opinion holder refers to the individual or organization that is expressing an opinion on a particular entity (B. Liu, 2012). In Figure 1.1, the opinion holder is the person who is stating the opinion. In the case of movie reviews for example, the opinion holder is the author of the particular review.

1.1.2.2 Object (Entity)

Opinions can be expressed over any given thing such as products, services, events, topics etc. by any individual or organization (Balazs & Velásquez, 2016). Seerat and Azam (2012) defined an object as a concrete or abstract item upon which an opinion is articulated. Generally, an object is referred to as the entity upon which an opinion has been expressed by the opinion holder. The book is the object on which the opinion holder (person) expresses the comment "*This is a great book*" in Figure 1.1

1.1.3 Regular and Comparative Opinions

Opinions can be categorized as regular (B. Liu, 2007) and comparative opinions (Jindal & Liu, 2006). There are two sub-types of regular opinions, namely, direct and indirect opinion. A direct opinion is expressed directly on an entity while an indirect opinion is conveyed indirectly (B. Liu, 2007). Examples of direct and indirect opinions are as follows:

Direct Opinion: "*The car has great horsepower*" - the positive sentiment on the horsepower of the car directly impacts the opinion of the car.

Indirect Opinion: "*After my recent service, my car seems to perform better*" - expresses a positive response of the service on the performance of the car.

A comparative opinion dictates a relation of resemblances or variances between two or more entities and/or a partiality of the opinion holder with respect to shared aspects of the entity (Jindal & Liu, 2006). For example, "*The crust from Dominos is better than Pizza Hut*" asserting the author prefers Dominos over Pizza Hut. Jindal and Liu (2006) discovered most comparative opinions use comparative or superlative forms of adjectives or adverbs although this may not always be the case. The following section will look into components of emotion analysis.

1.1 Emotion Analysis

Recent literature have started to look into emotion as part of sentiment analysis, and have defined emotion as a quintuple (e, a, m, f, t) where (e) represents target of the entity, (a) refers to target aspect of (e) that is accountable for the emotion, followed by (m) which is the type of emotion and its intensity, (f) is the entity that feels the emotion and (t) represents time the emotion was expressed (Cambria et al., 2017; Rosso, Bosco, Damiano, Patti, & Cambria, 2016; Yadollahi et al., 2017). As an example, take the emotion expressed in the text "*I am so thrilled with the football team captain today*"; (e) is represented by the football team, (a) is the team captain, the emotion is joy, and the use of the word *thrilled* shows its intensity is greater than *happy*. The feeler of the emotion (f) is the author of the statement and (t) is depicted as *today*.

In a survey conducted by Yadollahi et al. (2017), four distinct emotion analysis tasks were identified. The first task, emotion detection, is almost similar to subjectivity detection whereby the purpose is to detect and identify emotion that is found within a text (Desmet & Hoste, 2013). Results returned are similar to sentiment detection (positive or negative). For example, "*I can't believe my good fortune this morning*" shows positive detection for the emotion. The second task, emotion polarity classification, refers to the task of detecting the polarity or intensity of emotion within a text (Quan & Ren, 2010). For instance, "*The ending of the movie was very displeasing*" shows disgust but due to the use of the word *very*, the emotion detected within this sentence is of a higher polarity. The third task, emotion classification, is defined as a fine grained task of classifying the emotion detected into joy, sadness, trust, fear, disgust etc. (Jun Li, Rao, Jin, Chen, & Xiang, 2016) based on specific emotion frameworks, such as the Plutchik (2003) emotion wheel. For example, "*The outcome of the game tonight is really upsetting me!*" would classify this text as anger. The final task is emotion cause detection where the focus is to mine factors for eliciting some form of emotion (Gao, Xu, & Wang, 2015). As an example, "*The weather forecast shows its sunny tomorrow, I hope we can have a happy picnic*", this text corresponds the emotion hope with the prospective event of the picnic organized tomorrow.

1.2 Sentiment Analysis

In this section, a brief introduction to sentiment analysis is discussed. Sentiment analysis is primarily a study on opinions that indicate positive and negative sentiments. It is considered an active area of research and the continuous interest in this area is coupled with the expansion of social networking sites and global dependency on the Internet (Habernal, Ptáček, & Steinberger, 2014; Hays, Page, & Buhalis, 2013; Razzaq, Qamar, & Bilal, 2014).

1.2.1 Levels of Sentiment Analysis

Generally, sentiment analysis can be categorized into three levels; document, sentence and aspect (Medhat, Hassan, & Korashy, 2014). The purpose of document level sentiment analysis is to classify opinions of the document as either positive, negative or neutral where the assumption is that the whole document is discussing a single topic (Moraes, Valiati, & Neto, 2013).

The sentence level sentiment analysis on the other hand, works on classifying sentiment expressed in each individual sentence. A sentence is first identified as either an objective or subjective sentence where only subjective sentences are taken into consideration in classifying opinions as positive or negative (Medhat et al., 2014). Although Wilson, Wiebe, and Hoffmann (2005) has observed how sentiment expressions are not inevitably subjective in nature, literature has not discovered a vast difference between document and sentence level classification as sentences are simply regarded as short documents (B. Liu, 2012; Medhat et al., 2014). Furthermore, text classified at a document or sentence level does not cater to the need of opinion extraction on specific aspects of an entity which is crucial in many applications, as these levels of opinion extraction are more semantic centric (Jeyapriya & Selvi, 2015; Lek & Poo, 2013; Patra et al., 2014)

Aspect level sentiment analysis focuses on classifying sentiment with regard to specific aspects of an entity (Patra et al., 2014). Each entity and their aspects need to be identified before proceeding with this form of classification (Lek & Poo, 2013). An opinion holder would be able to give opinions on different aspects of the same entity (Medhat et al., 2014), for example, "*The quality of the picture and zoom is amazing on this camera, but the filter options are disappointing*", where the picture quality and zoom aspect of the camera (i.e. the entity) share a positive sentiment, but the filters are marked with a negative sentiment.

1.2.2 Sentiment Analysis Approaches

Medhat et al. (2014) identified two approaches in conducting sentiment analysis: lexicon-based and machine learning (Figure 1.2). The lexicon-based approach requires suitable lexicon construction while the machine learning approach automatically classifies text with the aid of training data sets derived from human annotation (Paltoglou & Thelwall, 2017).

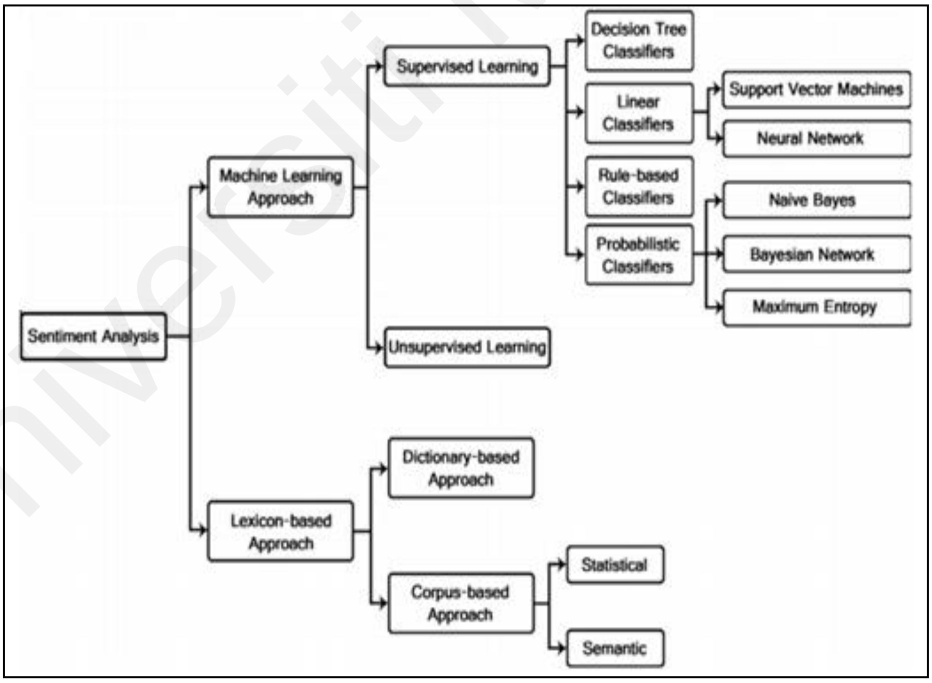


Figure 1.2 Sentiment analysis techniques (Medhat et al., 2014)

The lexicon-based approach is further divided into dictionary based and corpus-based approaches while the machine learning approach can be divided into supervised and unsupervised learning. The following sub-sections will briefly discuss the aforementioned approaches.

1.2.2.1 Dictionary Based Approach

A lexicon is a set of words that contains both positive and negative words with the corresponding sentiment score (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). In the dictionary-based approach, the lexicons are manually compiled as is the case of General Inquirer (Stone, Dunphy, Smith, & Ogilvie, 1968) which is still used till date (Mummalaneni, Gruss, Goldberg, Ehsani, & Abrahams, 2018). Dictionaries such as WordNet (Miller, 1995) used for sentiment analysis also provide antonyms and synonyms for each word. The dictionary-based approach provides an added leverage as one can effortlessly find a large number of sentiment words with their orientations without having to create one. Nevertheless, most sentiment analysis studies are domain dependent, and therefore employing a dictionary that is domain independent would compromise the sentiment accuracy (Bravo-Marquez, Frank, & Pfahringer, 2016; Mishra, Venugopalan, & Gupta, 2016).

1.2.2.2 Corpus Based Approach

The corpus-based approach employs two methods: the statistical method adapts a list of known sentiment words to discover other sentiment words using the same domain corpus (J. Zhao, Liu, & Xu, 2016), whereas the semantic approach uses syntactic relations of opinions and targets to extract sentiment words (Ristoski & Paulheim, 2016). Although corpus-based approaches are handy in detecting domain-specific words and their sentiment orientation, the approaches are restricted in terms of detecting contextual subjectivities and sentiments at sentence level (J. Zhao et al., 2016). In short, despite a word in the lexicon marked as positive or negative, when used within the context of a sentence, it may not carry any sentiment value (Song et al., 2016).

1.2.2.3 Supervised Learning

As shown in Figure 1.2, the machine learning approach is branched out as supervised and unsupervised learning. This approach depends on machine learning algorithms to classify text using syntactic and/or linguistic features (Medhat et al., 2014).

Supervised learning uses two sets of documents for training and testing purposes. The training set is used to train the algorithm on different feature characterization and use the trained algorithm to identify the same patterns for classification purposes based on the testing data set (Deng, Luo, & Yu, 2014). Naïve Bayes and Support Vector Machines are among two of the most effective classification algorithms (Medhat et al., 2014), however, with the recent diversity of dataset, other algorithms such as Decision Tree, Random Forest, Logistic Regression etc. have also emerged to be as effective (Balahur & Turchi, 2012; Duwairi & Qarqaz, 2014; F. H. Khan, Qamar, & Bashir, 2016a; Rohani & Shayaa, 2015). Nevertheless, one of the major concerns of adopting supervised learning is the availability of a proper set of training data as a very extensive chunk of training data is needed to train the algorithm and produce better accuracy. This is costly not only in terms of time but monetary as well (Medhat et al., 2014).

1.2.2.4 Unsupervised Learning

As mentioned in the section above, in order to classify data using supervised learning properly, a large number of labelled training data is needed. In some occasions, collecting such a large number of labelled data may be difficult therefore, it would be easier to adopt the unsupervised learning methods (Medhat et al., 2014).

Maas et al. (2011) found the classification technique adopted in unsupervised learning is based on fixed syntactic patterns composed of part-of-speech (POS) tagging where a word can be a noun, verb or adjective depending on the context the word was used. For example, the word “light” can be considered a noun (“*Switch on the light*”), a verb (“*Light this place up with candles*”) and an adjective (“*Please do not take this lightly*”). Hence, POS tagging is crucial to determine which sentiment the different forms of the word “light” belongs to. Nevertheless, a fully unsupervised learning method has a tendency of producing incoherent results due to the absence of training data, and thus rendering it impossible for human annotations (Maas et al., 2011).

1.2.2.5 Semi-Supervised Learning

The semi-supervised learning approach is a combination of both supervised and unsupervised learning methods. The biggest distinction between supervised and unsupervised approach lies within the dataset the algorithm is trained upon. Supervised learning algorithms are trained on labelled datasets that direct the algorithm to better understand important features used for classification purpose (Samal, Behera, & Panda, 2017; D. Vilares, Alonso, & Gómez-Rodríguez, 2017). Unsupervised machine learning algorithms however, are trained using unlabeled data leaving the algorithm to determine substantial features independently based on intrinsic patterns available within the data itself (da Silva, Coletta, Hruschka, & Hruschka Jr, 2016; Fernández-Gavilanes, Álvarez-López, Juncal-Martínez, Costa-Montenegro, & Javier González-Castaño, 2016; S. Lim, Tucker, & Kumara, 2017). The semi-supervised learning approach uses both supervised and unsupervised methods to train an algorithm (D. Anand & Naorem, 2016; F. H. Khan, Qamar, & Bashir, 2016b), and they are particularly useful when there is a limited amount of labelled data available (Charalampakis, Spathis, Kouslis, & Kermanidis, 2016; Krawczyk, Minku, Gama, Stefanowski, & Woźniak, 2017). This approach helps to eliminate the need of labelling large amounts of data (i.e. supervised learning approach) that is not only time consuming but costly as well (Charalampakis et al., 2016). Furthermore, extensive labelling may also lead to human biases within the model. Past studies adopting semi-supervised learning approaches have discovered that the accuracy of a classification model can be improved along with a reduction in cost and time in building a model by introducing unlabeled data amid the training process, (Altınel & Ganiz, 2016; Fernández-Gavilanes et al., 2016; Yingjie Tian, Zhang, & Liu, 2016).

1.2.3 Sentiment Intensity

Chaudhuri (2006) divides sentiment into rational and emotional sentiment, where rational sentiment comes from rational reasoning and tangible beliefs with no emotional expression. Emotional sentiment on the other hand, refers to emotional responses towards an entity which is a reflection of an individual's psychological state of mind. It has been found that emotional sentiment are much stronger compared to rational sentiment and are more important in practice (Cambria, 2016; Cambria et al., 2017). An emotional sentiment can have many emotions tied to it, e.g. anger, joy, fear, sadness etc.

Sentiment can have a diverse level of strength or intensity. There are two distinct manners used within a written text to convey the intensity of one's feelings (Cambria, 2016; H. Jiang, Qiang, & Lin, 2016; Haiqing Zhang, Sekhari, Ouzrout, & Bouras, 2016). The first method is the choice of words or phrases that portray different levels of strength. For example, choosing *excellent* over *good*, or *loathe* instead of *dislike*, where the former in both examples expresses a stronger sentiment compared to the latter. The other method uses intensifiers (used to increase positive/negative sentiment) or diminishers (used to decrease positive/negative sentiment). Examples of intensifiers include *very*, *so*, *terribly* etc. and examples of diminishers include *barely*, *somewhat*, *slightly* etc. In this research, the interest was to look upon other measures that can be used to indicate a stronger sentiment intensity which can contribute to a more accurate classification scheme.

1.3 Purpose Analysis

A literature study conducted by Yan, Wang, Chen, and Zhang (2016) found 23% of the world population referred to the Internet to look for health related information. In China alone, 64% (more than 100 million hits) of Internet users frequented social media sites looking for health related information, to share personal experiences as well as discuss medical treatment options on a monthly basis (Yan et al., 2016). Social media sites such as Facebook and Twitter are most widely used by patients and caregivers to cultivate social support which includes informational, emotional and appraisal support as well as providing a sense of companionship to one another (Abedin et al., 2017; Rus & Cameron, 2016; Sharma, Yadav, Yadav, & Ferdinand, 2017; Y. Zhang, He, & Sang, 2013). This is a clear indication on the role social media platforms play in supporting those seeking health related information in the current era. However, studies related to sharing of health-related information on social media platforms have highlighted one crucially missing link which is the need for users to search for relevant information with ease through the thousands of posts and texts available (AlQarni, Yunus, & Househ, 2016; McRoy et al., 2018; Salas-Zárate, Medina-Moreira, Lagos-Ortiz, Luna-Aveiga, Rodríguez-García, et al., 2017; Yan et al., 2016)

Literature has defined purpose as reason or intent behind a statement made (S. M. Mohammad, Zhu, Kiritchenko, & Martin, 2015). S. M. Mohammad et al. (2015) studied electoral tweets and classified them according to sentiment, emotion and purpose claiming it was necessary to discover the intent behind an emotion in order to achieve a better understanding on the emotion being displayed. This is because an emotion such as disgust can also be associated with the intent to either vent, ridicule or show disappointment. In the context of this research however, purpose analysis looks into classifying extracted posts according to pre-determined topics known as purpose (i.e. purpose of the text posted). The said post for example, could either be to disseminate information, provide update on latest treatment options or seeking emotional support from other users who are battling the same disease.

1.4 Problem Statement

The problem statements of this research are discussed within this sub-section. This research has identified three areas upon which a contribution can be made; introducing a multi-tier classification algorithm for a health-related dataset, using integrated Facebook features to calculate sentiment and emotion intensity and proposing a weighted semi-supervised classification algorithm using string vector with dimensionality reduction.

1.4.1 Lack of Studies Adopting Multi-Tier Classification Framework

The Merriam-Webster dictionary defined multi-tier as many series of rows or ranks, one raising above another. Each tier lies independent of each other with little or no connection between each other. However, in the context of this research, data from one tier will be provided as an input to the next tier creating the illusion of a hierarchical structure yet the processing of each tier is distinct one from another.

The expansion of social media has encouraged the mushrooming of online health related communities that are a major source of support for people with health problems (K. Zhao et al., 2014). With the ever-growing repository of information available, it is an arduous task for users to be able to filter the relevant needed information from the massive set of unstructured data available. Therefore, it is important to incorporate a hierarchical structure that would be able to classify the unstructured data according to proper classification themes thus improving classification accuracy.

From a patient's perspective for an example, there are some who are more concerned with finding out the type of medication that could help battle a type of disorder, and how some patients are prescribed modern medication whilst others are advised to change their lifestyles. Classifying each item in its relevant theme will also help patients to navigate themselves easily through the large pool of data to find the information needed. For example, a simple search of "*alternative throat cancer treatment*" may return thousands of results that need to be filtered. This may cause stress as well as confusion to the patient thus rendering them helpless. The current state of information available on such platforms requires patients to rummage through thousands of unstructured data. Therefore, by proposing a multi-tier classification framework, the information could be sorted and displayed in such a manner that eases information retrieval for patients (Du, Liu, Ke, & Gong, 2018) thus creating a tier-like structure of information. A hierarchical tree would provide a road map to users in getting the exact information with ease, in a more effective and efficient manner.

Consequently, from the framework perspective, a tree-like hierarchical structure not only helps the algorithm to classify sentiment more accurately, and thus improving the performance of the classifier, but also eliminates irrelevant examples by upper-level classifiers, hence making it easier for lower-levels to classify data (Xu, Yang, & Wang, 2015). Figure 1.3 is a simple illustration of what information in a multi-tier framework would look like using the cancer treatment example stated above.

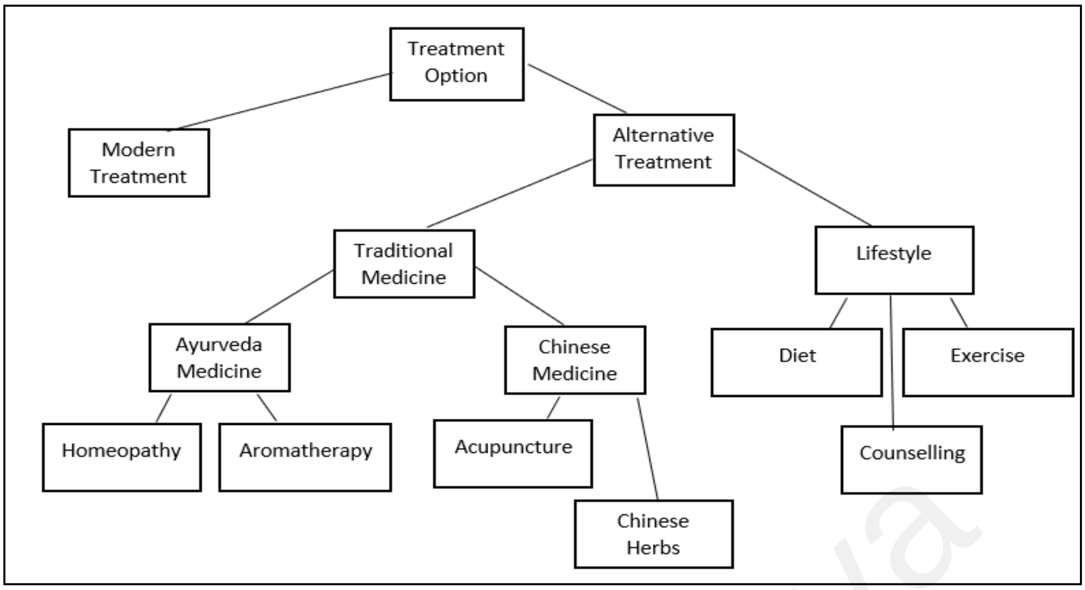


Figure 1.3 Multi-Tier Classification of Cancer Treatment

Previous studies adopting the multi-tier framework only looked into classifying sentiment into different levels. Moh, Gajjala, Gangireddy, and Moh (2015) classified movie reviews in a tree-like manner where the parent node was positive followed by two children nodes (less positive, more positive), likewise for negative reviews. Jinyan Li, Fong, Zhuang, and Khoury (2016) classified online news sentiment by applying a multi-tier filtering on the training dataset to test factors that influence the performance of the classification algorithm. To-date, there is no work related to classifying health related dataset in a hierarchical manner while incorporating different elements within the same tree like structure.

1.4.2 Using Facebook Features for Facebook Sentiment and Emotion Analysis

Communication on Facebook is not only limited to posting and commenting but also includes sharing, liking and reacting. With over 1.32 billion active users worldwide and an average of 4.75 billion items being shared on a daily basis, Facebook has been regarded as a mega warehouse of data (P.-W. Fu, Wu, & Cho, 2017). Statista[1] has ranked Facebook as the most popular social networking site followed by YouTube, WhatsApp and WeChat. Despite the generous word count limit on Facebook, Pool and Nissim (2016) found that posts tend to receive more *like* or *share* compared to longer comments. Furthermore, the Facebook EdgeRank[2] algorithm assigns different weights on each behavior (*like*, *comment*, and *share*) where the highest weight is assigned to *share* and least to *like* (C. Kim & Yang, 2017). Facebook introduced the *reaction* feature in 2016 which entails five pre-defined emotions with the corresponding emoji (love, haha, wow, sad, angry), permitting users to express their feelings wordlessly with just a click of a button (Smieško, 2016).

Various studies on Facebook features have been carried out in the past (C. Kim & Yang, 2017; Oeldorf-Hirsch & Sundar, 2015; Pool & Nissim, 2016; Smieško, 2016). Oeldorf-Hirsch and Sundar (2015) studied how the incorporation of such features facilitated discussion of news and found the more users engaged in discussion via comments, the higher the presumption of authenticity of the news. Smieško (2016) alternatively studied how the usage of Facebook *reaction* contributed to hate speech in Slovak Republic. Studies have also found that the use of such nuances led to showcasing preference, for example, a higher number of likes and comments indicates users are in agreement with the content (Oeldorf-Hirsch & Sundar, 2015; Zell & Moeller, 2018), or an increase in the number of share demonstrates a higher importance of an item to be known by all, hence increasing its visibility and influence (Carah, 2014; Coursaris, van Osch, & Balogh, 2016; C. Kim & Yang, 2017). Nevertheless, a study leveraging on all

[1]https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/
[2]http://edgerank.net/#How-does-EdgeRank-work

these users' behavior to attain a higher accuracy for sentiment has not been carried out to-date. This study believes the use of such features may increase the intensity of the sentiment being displayed, hence improving their scores. This research is interested in exploring this avenue to prove this hypothesis.

1.4.3 Weighted Classification Algorithm using String Vector

Text classification is a task of assigning predefined categories to unseen documents (Mirończuk & Protasiewicz, 2018). Traditionally, text classification documents are encoded in numerical vectors where a simple binary value of whether a word exists or not, is generally used. The vector space model is the most frequently used model for classification where each document is represented as a vector (Jagtap & Adamuthe, 2018; Mirończuk & Protasiewicz, 2018). According to Du et al. (2018), classes are categorized based on word patterns. For example, documents in class Government would have high values for words such as ministry, treasury and policy whereas documents in the class Science would show high values for words like physics, blackhole and anatomy. Therefore, documents in both classes form distinctive adjacent classes as shown in Figure 1.4 which can then be used to classify new documents. However, the use of vector space models lead to problems such as huge dimensionality (Al-Anzi & AbuZeina, 2017; Y.-S. Lin, Jiang, & Lee, 2014) and sparse distribution (Guo, Shi, & Tu, 2016; Williams & Gong, 2014)

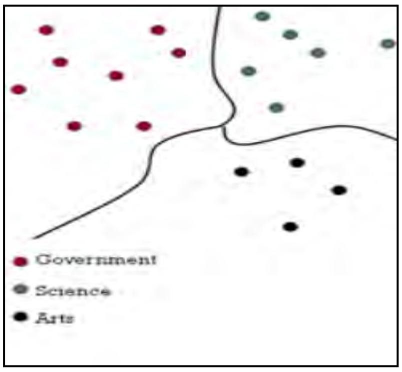


Figure 1.4 Vector Space Classification into three classes (Mirończuk & Protasiewicz, 2018)

The problem with huge dimensionality persists due to the large number of features required to robustly represent the document as a numerical vector (Mirończuk & Protasiewicz, 2018). In such cases, the number of features identified within the text is much larger than the number of labelled data itself (Al-Anzi & AbuZeina, 2017). This proves to be costly as labelled data for text classification needs to be prepared manually, hence identifying a large number of features to ensure accurate classification will be a time-consuming process.

Numerous text classification studies have been carried out in the past (Al-Anzi & AbuZeina, 2017; Guo et al., 2016; Y.-S. Lin et al., 2014). For example, Al-Anzi and AbuZeina (2017) used a singular value decomposition method to extract features based on latent semantic indexing (LSI) and found LSI to be a better textual representation technique as it maintains the semantic information between words. Jo (2017a) used string vector with k-nearest neighbor algorithm, with results showing an improved accuracy of text classification by 5% compared to using the traditional method of numerical vectors. In another work, Jo (2017b) used string vectors to modify the Agglomerate Hierarchical Clustering (AHC) algorithm and discovered the algorithm was able to classify text more transparently. This shows converting numerical vectors to string vectors improves dimensionality reduction, hence improving classification accuracy.

This research proposes to assigned a computational weight using a feature selection algorithm and convert the classification algorithm using string vectors in order to reduce the dimensionality, and classify text more accurately.

1.5 Research Aim

The aim of this research is to automatically classify sentiment, type, emotion and purpose, using diabetes dataset gathered from Facebook. Literature shows classification accuracy increases as text gets classified into more specific groups moving down the tiers of a multi-tier framework (Xu et al., 2015). Therefore, this research is looking to use this information to experiment on adopting a multi-tier framework that will be able to classify text for sentiment, type, emotion and purpose within a single framework, and thus providing more accurate results. This will help online users to obtain the needed relevant information they are seeking without scavenging through thousands of data (Salas-Zárate, 2017).

The focus of this research is to improve classification so that it would be able to almost mimic how a human classifies sentiment, type, emotion and purpose, given a text. In order to achieve this, features that have shown to have impact on the way humans interact, yet have not been used in sentiment and emotion analysis studies will be taken into consideration (C. Kim & Yang, 2017).

It is the aim of this study to investigate and experiment on the extracted dataset in order to better classify text for all four elements of sentiment, type, emotion and purpose so that future studies would be able to use this research work and further improve the classification process.

1.6 Research Objective and Questions

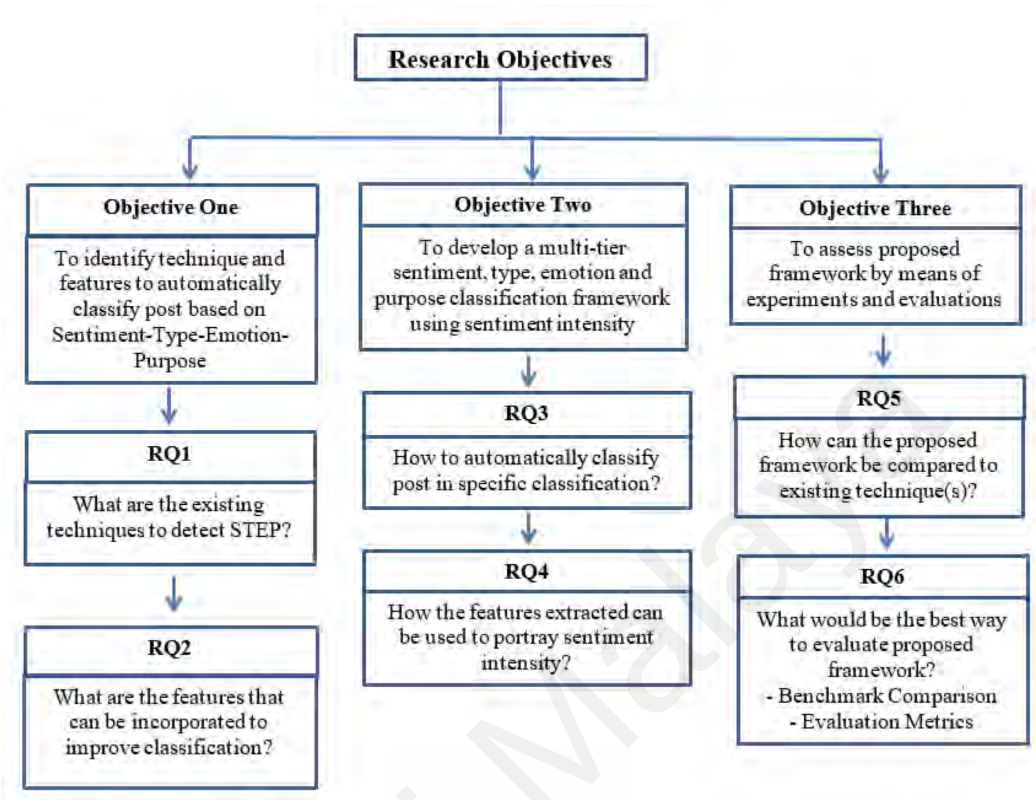


Figure 1.5 Research Objectives and Research Questions

Figure 1.5 displays the research objectives and their corresponding research questions identified from the problem statements above. A detailed discussion of the three research objectives and research questions are as follows:

First objective: To identify techniques and features to automatically classify posts based on Sentiment, Type, Emotion and Purpose (STEP).

The first aspiration of this research is to identify techniques and features that are currently being used to classify posts based on sentiment, type, emotion and purpose. The idea is to find pre-existing techniques that have been identified to classify multiple elements (S. M. Mohammad et al., 2015), and to identify techniques that have shown promising results (Baqapuri, Saleh, Ilyas, Khan, & Qamar, 2016; Ive, Gkotsis, Dutta, Stewart, & Velupillai, 2018; Jagtap & Adamuthe, 2018; Moh et al., 2015; Mujtaba et al., 2017). Once the techniques have been identified, this research will look into other possible features such as number of likes, comments, shares and reactions to improve the classification techniques (W. Kaur, Balakrishnan, Rana, & Sinniah, 2019).

Second objective: To develop a multi-tier sentiment, type, emotion and purpose (STEP) classification framework using sentiment intensity.

The objective of this research is to classify posts extracted from Facebook diabetes community according to sentiment, type, emotion and purpose (STEP). With the techniques that have been identified from the first objective, this research will then progress to automatically classify posts with respect to sentiment, type, emotion and purpose (STEP) by taking sentiment intensity into consideration. The intensity will be determined based on the identified features. A detailed description of the framework and how the sentiment intensity was determined will be discussed in Chapter 3.

Third objective: To assess the proposed framework by means of experiments and evaluations.

The proposed classification framework will be evaluated using the standard evaluation metrics of accuracy, area under curve (AUC) and F1-Score (Ravi & Ravi, 2015). Each tier will be tested individually as well as the framework as a whole against existing benchmark models.

1.6.1 Research Questions

(a) ***RQ1. What are the existing techniques to detect Sentiment-Type-Emotion-Purpose (STEP)?***

This research question explores possible technique(s) that are currently being used to detect sentiment, type, emotion and purpose. This is to build an understanding and base foundation on how classification process for all four elements are being carried out, and to investigate the gaps in current technique(s) that can be offered as a contribution to this research.

(b) ***RQ2. What are the features that can be incorporated to improve classification?***

This is also an exploratory question looking into possible features such as the number of shares, comments, likes and reactions that can be incorporated to further improve sentiment classification process compared to the pre-existing features that have been used in other studies (Bilici & Saygın, 2017; C. Kim & Yang, 2017; Meire, Ballings, & Van den Poel, 2016; Quesenberry & Coolsen, 2018; Zell & Moeller, 2018). Other features that can be included to improve classification process are also explored to help increase classification accuracy for the other tiers as well (type, purpose and emotion).

(c) ***RQ3. How to automatically classify posts in specific classifications?***

The second objective of this research is to be able to automatically classify online posts for sentiment, type, emotion and purpose. Since all four elements are distinct from one another, it is necessary to explore options to classify the said posts automatically according to each element. Therefore, this question will look into current ways posts are being automatically classified among all elements individually before proposing a manner to classify them within a single framework.

(d) ***RQ4. How the features extracted can be used to portray the opinion strength?***

With respect to the second research question, once features that contribute towards classification have been identified, this research will investigate mechanisms to convert those features to measurable means that can show an increase or decrease in the opinion strength.

(e) ***RQ5. How can the proposed framework be compared with existing framework(s)?***

The third objective looks into ways of experimenting and evaluating the proposed framework. This research question prompts this research to conduct experimentations comparing the proposed framework against other framework(s) that have been used to classify sentiment, type, emotion and purpose for evaluation.

(f) ***RQ6. What are the metrics that can be used to evaluate the proposed framework?***

The standard evaluation metrics (accuracy, area under curve and F1-Score) will be applied when evaluating the proposed framework. The results will also be compared with existing technique(s) to analyze the performance of the proposed framework. Apart from that, a comparison between the proposed framework and benchmark models will be conducted to evaluate the ability of the framework to produce results as close as possible to human annotation.

1.7 Research scope

A brief description of the scope and limitations of this research will be defined within this section. This is crucial in order to set clear goals aligned to the research objectives.

The dataset used for this research will be extracted from social media platforms, particularly Facebook. Users are progressively turning towards social media platforms for health information as well as sharing health care experiences (Martínez et al., 2016; Rodrigues, das Dores, Camilo-Junior, & Rosa, 2016). Apart from Wikipedia and online health forums, Facebook is one of the most active social media platforms hosting over half a million health related groups and pages; some of which have no less than thousand active users (Lu, Wu, Liu, Li, & Zhang, 2017). With dedicated groups and pages towards discussing and sharing information to specific diseases, it is easier to collect data that belong to the same sub-group instead of a mixture of other diseases as is the case on some online forums (Korkontzelos et al., 2016). Hence, for this research, data are extracted from Facebook groups and pages.

The number of pages and groups on Facebook related to health or disease is abundant, therefore, to narrow the scope and test the proposed framework, it is important to choose one of the many diseases available. There are some groups and pages related to health that are purposely kept private such as cancer and mental health groups. Diabetes was once known as adult-onset diabetes, however that presumption proves to be premature in today's context as even young children are diagnosed with it (Association, 2018). Today, diabetes has become a pandemic and it is necessary for healthcare systems as well as pharmaceutical companies to develop new methods to enlighten the general public on the ramification of this disease outside from the clinical or offline setting (Salas-Zárate et al., 2017). Even closer to home, the National Health and Morbidity survey 2015 has stated one in five Malaysians are affected by this disease. Considering the chronic nature of the disease, the current research aims to focus on the Facebook diabetes community.

Possible limitations of this research and scope of this research will be further elaborated in Chapter 5.

1.8 Significant research contributions

The contributions of this research are as discussed below:

a) The first contribution is to use string vectors instead of numeric vectors. Opting for numeric vectors have caused huge dimensionality issues in classifying text correctly. Previous studies have looked at opting for diverse preprocessing techniques in order to conquer this problem (Altınel, Can Ganiz, & Diri, 2017;

Diab & El Hindi, 2017; Ive et al., 2018; Mirończuk & Protasiewicz, 2018; Onan, Korukoğlu, & Bulut, 2017). This research aims to address this issue by converting the numeric vectors to string vectors with weighted features.

b) The second contribution is extracting features that can be used to measure sentiment intensity and manipulating those features to affect the final sentiment score. A novel measurement technique using those features as an added weight for sentiment classification is introduced in this research. Each feature (i.e. like, comment, share and reaction) is assigned a different weight and is mathematically added to the final sentiment score, resulting in a more accurate result. Previous studies have only looked into studying the behavior of the extracted features (Carah, 2014; C. Kim & Yang, 2017; Quesenberry & Coolsen, 2018), without incorporating them to improve sentiment classifications.

c) The final contribution is in the form of a multi-tier framework that combines four elements; sentiment-type-emotion-purpose. The significance in having such a framework lies in the ability of each tier to process the data individually and pass of relevant data to the next tiers. This helps eliminate unnecessary data that does not make a contribution towards the classification process (Baqapuri et al., 2016; Jagtap & Adamuthe, 2018; Kowsari et al., 2017)

1.9 Research Methodology

Figure 1.6 is a general depiction of the methodology adopted in this research. This study also adopted the following steps:

i. Conducting a literature review to identify techniques used in classifying sentiment, type, emotion and purpose

ii. Conducting a literature review to identify features that can be incorporated to improve classification

iii. Problem identification from literature review

iv. Conducting interviews with medical professionals, patients and caregivers to help define classes related to treatment, medication etc.

v. Data collection from diabetes related Facebook pages

vi. Conducting human annotation survey

vii. Developing STEP framework

viii. Evaluating proposed framework with respect to evaluation metrics and benchmark model comparisons

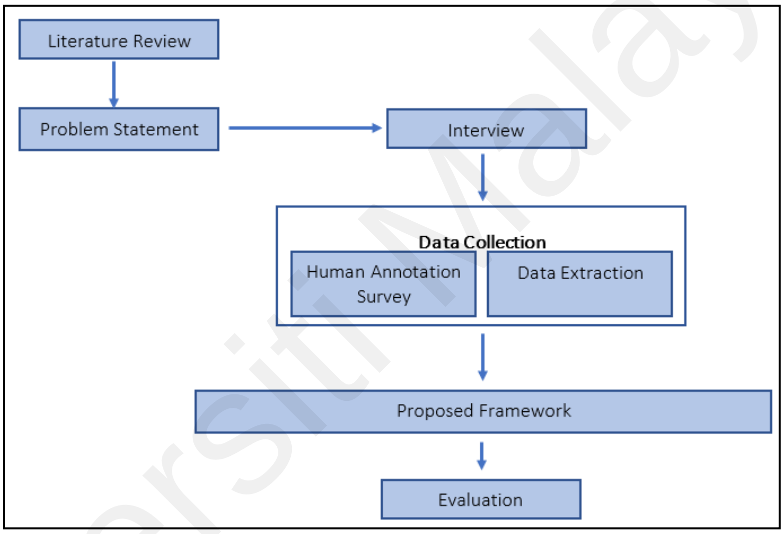


Figure 1.6 General overview of methodology

1.10 Thesis layout

This thesis consists of five chapters, and they are organized as per below:

i. Chapter 1: Introduction

This chapter serves as an introductory for sentiment analysis (or opinion mining) technique. The importance and other information of this technique has been clearly explained. Besides, it presents the problem statements, objectives, research questions and project scope-contribution identified for this research.

ii. **Chapter 2: Literature review**

This chapter discusses the review of literature on the relevant concept of research which comprises of current knowledge in addition to reviewing substantive findings towards the topic of research. This includes a review on previous studies adopting a multi-tier framework as well as Facebook behaviors and other semi-supervised machine learning algorithms. An overview of opinion mining and a detailed literature study on sentiment, emotion and purpose will also be discussed in this chapter.

iii. **Chapter 3: Research methodology**

This chapter provides the detailed overview of the research plan. It discusses the different stages involved in this research, including design of the data collection process, data cleaning, and data analysis. It also presents a detailed discussion of every novel contribution of this research which encompasses the formula proposed to calculate sentiment intensity based on Facebook features and the multi-tier framework of classifying sentiment, type, emotion and purpose.

iv. **Chapter 4: Results and discussion**

This chapter will discuss the results of the evaluation of the proposed framework by performing comparison between proposed framework and other frameworks in terms of accuracy, area under curve and F1-Score.

v. **Chapter 5: Conclusion, limitation and future work.**

This chapter will conclude the research, highlight the limitations and future study that would be helpful to improve the current framework.

CHAPTER 2: LITERATURE REVIEW

Chapter 1 focused on introducing basic concepts of the research including discussing the research objectives, scope as well as contributions. The aim of this chapter would be an in-depth study of the current literature work providing a foundation for this research. Figure 2.1 shows the roadmap for this chapter for ease of readability.

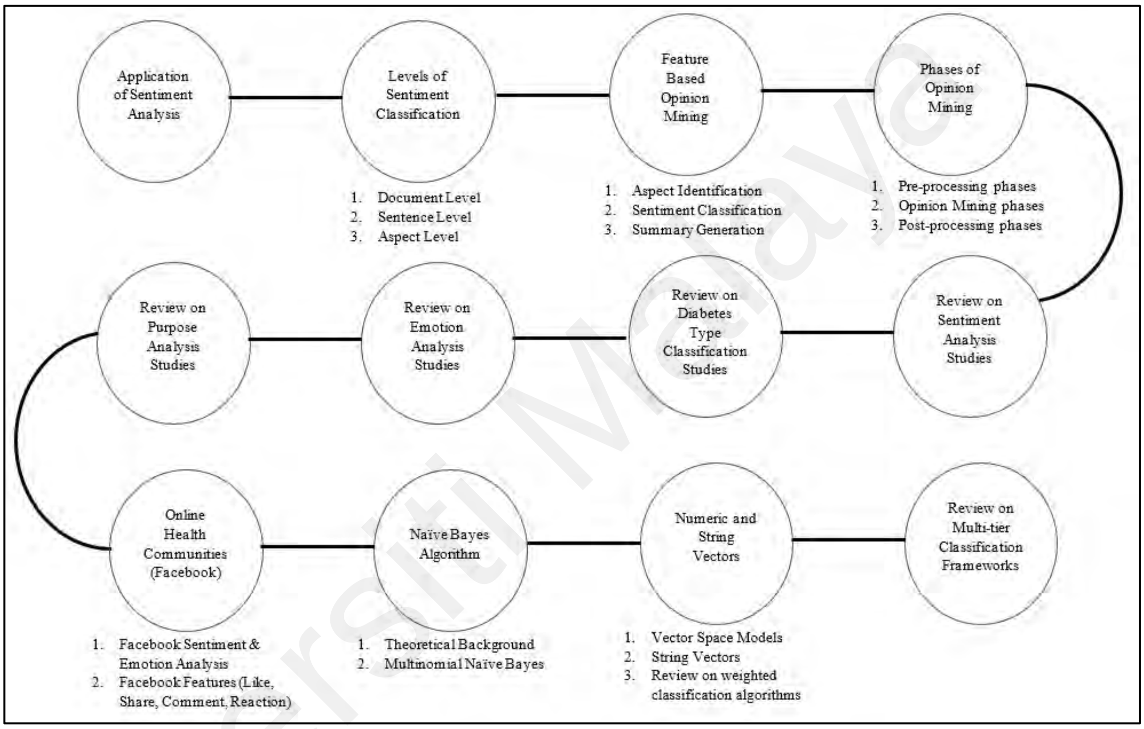


Figure 2.1 Chapter Roadmap

The chapter begins with an overview of the application of sentiment analysis, followed by a detailed discussion on levels of sentiment classification, feature-based opinion mining, phases of opinion mining, evaluation measurements, a discussion on the current work in the fields of sentiment and emotion analysis, followed by the multi-tier classification techniques, sentiment and emotion mining within the diabetes domain and finally inclusion of Facebook behaviors for sentiment analysis.

2.1 Application of Sentiment Analysis

As mentioned in Chapter 1, the application of sentiment analysis research today is not only limited to the business domain (H. Malik & Shakshuki, 2016; Teh, Pak, Rayson, & Piao, 2015; Vidya, Fanany, & Budi, 2015) but also expands itself into the entertainment (Hodeghatta, 2013; Nagamma, Pruthvi, Nisha, & Shwetha, 2015), political (Chang, Chiu, & Hsu, 2017; Charalampakis et al., 2016; Hammami, 2016), health (Ji, Chun, & Geller, 2013; Rodrigues et al., 2016; Wu, Moh, & Khuri, 2015) and education (Altrabsheh, Cocea, & Fallahkhair, 2014; Ortigosa, Martín, & Carro, 2014) domains, among others.

Products and services providers are interested in comprehending consumers' preferences and needs in a more accurate manner so that they are able to understand the driving force behind their decision to purchase a product/service (Jeyapriya & Selvi, 2015; Okada, Takeuchi, & Hashimoto, 2014). Consumers on the other hand are concerned with making informed decisions before buying a product or service (Das & Prathima, 2016). This form of mutual understanding between providers and consumers translates to better products and services being provided, which ultimately transcends towards a better quality of life for consumers and higher sale revenues for businesses (H. Malik & Shakshuki, 2016).

Similarly, in the political domain, sentiment analysis replaces the traditional method of opinion polls with the rapid and efficient automatic analysis of user contributed content leading to democracy in real-time (Chang et al., 2017; H.-C. Lin, 2017; Pak, Kim, Song, & Kim, 2015; Razzaq et al., 2014). Chang et al. (2017) states the ability of sentiment analysis and opinion mining lies within the numbers and level of detail that can be extracted from written text. Therefore, opinion polls are affordable when it comes to extracting major issues (Razzaq et al., 2014), however, sentiment analysis algorithms are able to draw out other information that can be deemed useful specifically enabling election candidates and even government administration to perceive public support when in engaging the public perspective on issues at hand (Vepsäläinen, Li, & Suomi, 2017).

The education sector has also turned towards conducting sentiment analysis studies as it has been discovered by Ortigosa et al. (2014), that such information helps schools understand the struggles students are facing and consequently, able to provide them with better suggestions in terms of course or subject selection. In fact, sentiment analysis studies have also been conducted for evaluation purpose on teaching staff, school facilities as well as course structure (Z. Liu et al., 2016). This is done in replacement of the traditional questionnaires or online survey forms that are much more structured but at times not really helpful especially in truly understanding how students feel.

News articles undoubtedly have a huge effect on our daily lives. Pröllochs, Feuerriegel, and Neumann (2015) revealed a negative financial news update has a tendency to send the stock market into a frenzy, hence reiterating the fact that the publishing financial news articles has a direct impact on a country's economy. Additionally, by being able to measure sentiment accrued from stock market news articles, Heston and Sinha (2014) were able to forecast stock market prices. . These findings are crucial specifically for government bodies that would be able to better manage possible economic threats to the country and nib them early (B. S. Kumar & Ravi, 2016).

Among other applications of sentiment analysis are improving recommender systems with the help of sentiment analysis techniques (Ghorpade & Ragha, 2012; Sundermann, Domingues, Conrado, & Rezende, 2016). W. Wang and Wang (2015) found that by analyzing sentiment of like-minded users of a product, their proposed recommender system was able to produce more accurate results. In another domain, the Indian government used sentiment analysis of tweets to track the Chikungunya and Dengue epidemic across New Delhi (Swain & Seeja, 2017). This goes to show the scope of contribution of sentiment analysis research is no longer tied to recommendations, user reviews or politics, but it has expanded itself towards other areas as well (Cambria, Schuller, Xia, & Havasi, 2013).

2.2 Levels of Sentiment Classification

As mentioned in Chapter 1, sentiment classification is carried out at three distinct levels, namely document level, sentence level and aspect level. This section provides a detailed explanation of all three levels of classification.

2.2.1 Document Level Classification

When classifying sentiment at the document level, the opinion is classified as positive, negative or neutral as a whole (A. Kumar, Kansal, & Ekbal, 2015). For example, in a case of movie reviews dataset, the reviews are classified as positive, negative or neutral based on the presence of opinion words within the text.

B. Liu (2015) has defined document level classification as a quintuple

(O, GENERAL, s, h, t),

Where given an opinion document D to evaluate the target object (O), to determine the sentiment (s) of the opinion holder (h) about the object (O), i.e. to detect (s) signified on the object GENERAL.

One of the concerns of this level is that a whole review is expressed as a single subject with an assumption that each opinion document contains opinion from a single opinion holder (B. Liu, 2015). Therefore, in cases of multiple subjects, this form of classification is unable to garner convincing results. Among studies that have adopted this form of classification in recent times include A. Kumar et al. (2015) who adopted the active learning technique in classifying tweets and Moraes et al. (2013) who compared the performance of Support Vector Machine algorithm against Artificial Neural Network when classifying movies dataset.

2.2.2 Sentence Level Classification

The objective of classification at this level is to segregate opinions into positive, negative and neutral at the sentence level. Each sentence is individually examined, thus chances of subjectivity classification during the pre-processing phase may occur (Gezici, Yanikoglu, Tapucu, & Saygin, 2012). Subjectivity classification refers to the process of grouping sentences into objective and subjective sentences, where objective sentences denote factual information and subjective sentences represent subjective opinions (B. Liu, 2015). As a result, there are two important tasks in conducting a sentence level classification:

Task 1: To identify if sentence contains opinions (subjective)

Task 2: To determine polarity of given sentence (positive or negative)

Recent works using this form of classification include Shoukry and Rafea (2012) who classified Arabic tweets and B. Yang and Cardie (2014) who proposed a context aware model for classifying customer reviews according to individual sentences.

2.2.3 Aspect Level Classification

Aspect level classification, also known as feature-based classification looks directly at the opinion instead of focusing on language constructions such as clauses, sentences or paragraphs (Ganeshbhai & Shah, 2015). Therefore, in the following literature, the term aspect level and feature-based will be used interchangeably referring to the fundamental concept of classifying opinions with respect to features. A feature-based classification is based upon the fact that users are known to articulate their opinion with reference to an entity rather than the entity itself (Z. Liu et al., 2016). For example, a feature-based classification on a camera would refer to opinion extraction based on the features of the camera (picture quality, battery life etc.) compared to talking about the camera itself.

Ganeshbhai and Shah (2015) states, a feature upon which an opinion is expressed is referred to as target of an opinion and without the identification of an opinion target, the opinion is of restricted use. This goes to show the importance of opinion targets in this form of classification, as in most cases a single sentence may contain multiple targets whereby each distinct target has its own set of opinions (Akhtar, Gupta, Ekbal, & Bhattacharyya, 2017). For instance, "Although the service was slow, the food was delicious"; here, it is evident that the food is assigned a positive sentiment however, the whole sentence cannot be considered positive.

The objective of a feature-based classification is to identify opinion targets represented by entities and/or their distinct aspects where a methodical summary of opinions about entities and their aspects are constructed (Ganeshbhai & Shah, 2015). Consequently, unstructured data are converted into structured data, which can then be used for both quantitative and qualitative analyses.

2.3 Feature-based Opinion Mining

This section takes an in-depth look into feature-based opinion mining and what distinguishes it from other forms of opinion mining. When it comes to opinion mining, an opinion of a single individual is not sufficient, therefore, opinion mining studies always look at extracting data from a large pool of opinions (Cambria et al., 2017). Typically, a feature-based opinion mining study comprises of three major tasks, namely aspect identification, sentiment classification and summary generation (Ganeshbhai & Shah, 2015). Let's take the following sentence as an example:

Had the best pasta yesterday, though it was a little on the pricey end.

From the above sample, it can clearly be seen how the opinion holder expresses a positive opinion on one feature (i.e. *pasta*) but a negative opinion on the *price* point of the *pasta*.

In his thesis, Gulaty (2016) described a feature-based opinion mining model as an object *O* with a finite set of features, $F = (F_1, F_2, \ldots, F_n)$ where each feature $f_i \in F$ is signified with a finite set of opinion words or phrases, $W = (W_1, W_2, \ldots, W_n)$ for the total number of features (n) of the object *O*. Likewise, an opinion document can be depicted with respect to a feature-based opinion mining model where an opinionated document D holds opinions on a set of objects $(O_1, O_2, \ldots, O_n)$ from a group of opinion holders $(h_1, h_2, \ldots, h_p)$.

The following subsections focus on the three key tasks of conducting a feature-based classification, that is, aspect identification, sentiment classification and summary generation.

2.3.1 Aspect Identification

The purpose of this task is to detect and extract related features from the text for summarization. Ganeshbhai and Shah (2015) defines the extraction of features with respect to the opinion tied to it to be the first step towards a feature-based opinion mining classification. Take for example the opinion depicted in Figure 2.2. The features identified are *food*, *portion* and *location.*

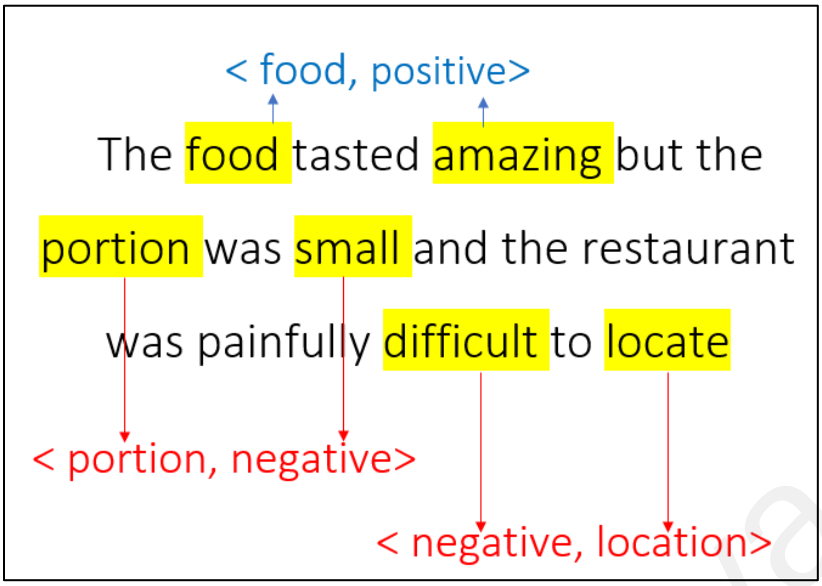


Figure 2.2 Identification of features

2.3.2 Sentiment Classification

Once the aspects or features have been identified, it is necessary to determine the sentiment (positive, negative, neutral) associated to each feature (Ganeshbhai & Shah, 2015). As it can be seen in the example shown in Figure 2.2, the sentiment associated to *food* is positive, meanwhile the sentiment for *portion* and *location* is shown as negative.

2.3.3 Summary Generation

The final task of feature-based opinion mining is to portray the processed results in a proper manner, which can easily be inferred by other users. Over the years, researchers looking into feature-based opinion mining have used pie charts, bar and line graphs for summarizing results to show most and least preferred features (Gulaty, 2016; Lek & Poo, 2013; Y. Zhao, Dong, & Yang, 2015), however, researchers are now opting to use various forms of advanced visualization techniques (Divya, Sandhya, & Sai, 2015; T. D. Nguyen, Tran, Phung, & Venkatesh, 2016).

2.4 Phases of Opinion Mining

Opinion mining adopts natural language processing techniques in classifying opinions and sentiments (S. Sun, Luo, & Chen, 2017). A review of the literature revealed three distinct phases in processing text for the purpose of extracting opinions, namely pre-processing, opinion mining and post processing, as depicted in Figure 2.3 (Singh & Kumari, 2016; S. Sun et al., 2017). The following sub-sections explains in detail each of the phases.

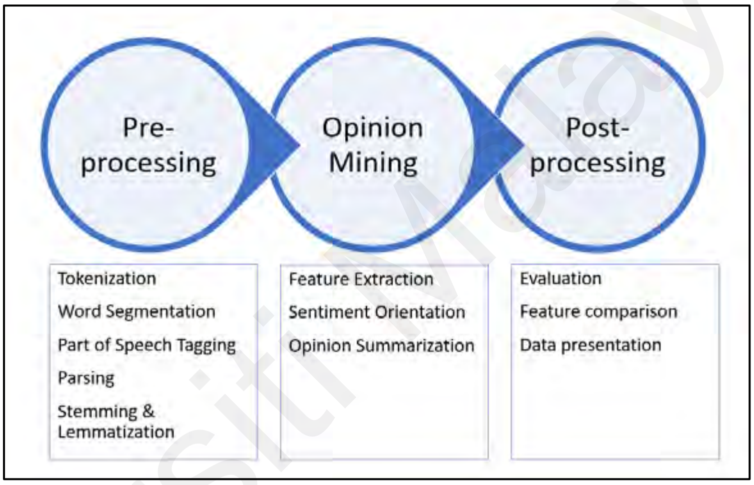


Figure 2.3 Phases of opinion mining

2.4.1 Pre-processing Phase

As shown in Figure 2.2, there are several tasks involved in the pre-processing phase of opinion mining, including tokenization, word segmentation, Part of Speech (POS) tagging, parsing, stemming and lemmatization (S. Sun et al., 2017). A brief explanation on the steps involved within the pre-processing phase is as below:

2.4.1.1 Tokenization

Tokenization is known as the basic building block for most natural language processing (NLP) task. A tokenizer works by splitting documents or sentences into tokens known as words or phrases. Simply stated, it splits individual sentences into words based on the space found within them (S. Sun et al. (2017), however, in the case of opinion mining, additional knowledge needs to be taken into account, such as opinion phrases and named entities (Singh & Kumari, 2016). It is also common for stop words such as “the”, and “a” to be removed at this stage as they do not contribute towards opinion extraction (Ghorpade & Ragha, 2012). There are several available tokenization tools freely available such as Stanford Tokenizer1[3], and OpenNLP Tokenizer[4].

2.4.1.2 Word Segmentation

Word segmentation is a serial classification problem used to classify opinions extracted from languages that do not have obvious word boundary markers, such as Chinese, Japanese and Vietnamese (S. Sun et al., 2017). In recent studies, word embedding and deep learning approaches have been adopted for word segmentation purpose (Ma & Hinrichs, 2015; D. Q. Nguyen, Vu, Nguyen, Dras, & Johnson, 2017). Freely available tools such as ICTCLAS[5], THULAC[6] and Stanford Segmenter[7] are also available for word segmentation.

2.4.1.3 Part of Speech Tagging

Part of speech (POS) tagging is responsible for examining the lexical information of a given text (Pirrelli & Zarghili, 2017). POS tagging works by associating part of speech

---

[3]http://nlp.stanford.edu/software/tokenizer.shtml
[4]https://opennlp.apache.org/documentation/manual/opennlp.html#tools. tokenizer
[5]http://ictclas.nlpir.org/
[6]http://thulac.thunlp.org/
[7]https://nlp.stanford.edu/software/segmenter.shtml

with each tokenized form of a text. In other words, each tokenized word is paired with its appropriate grammar tag according to the English language, such as noun, verb, adjective, adverb, pronouns, preposition (G. Wang, Zhang, Sun, Yang, & Larson, 2015) etc. Figure 2.4 is a pictorial depiction of POS tagging where each word is associated to a grammar rule. Such association is necessary because adjectives are opinion words and nouns are opinion targets (i.e. entities and aspects) (Sun et al., 2017). Therefore, using POS tagging is crucial to identify both opinions and opinion targets in opinion mining studies.

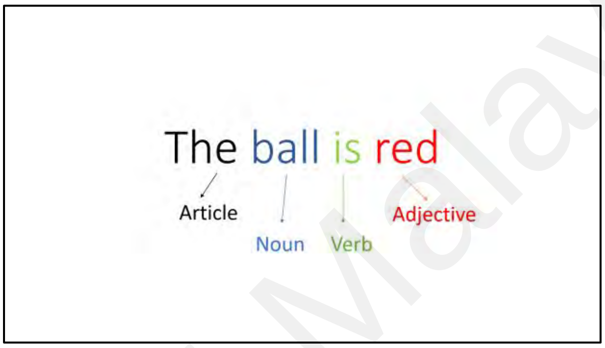


Figure 2.4 POS Tagging

2.4.1.4 Parsing

Unlike POS tagging that obtains lexical information, parsing acquires the syntactic information of a text (S. Sun et al., 2017). Pirrelli and Zarghili (2017) defined parsing as a tree-like structure that depicts the grammatical construction of a given sentence showcasing the correlation of different grammatical components. Therefore, compared to POS tagging, parsing provides a more elaborate information. Figure 2.5 shows a sample of parsing tree for the text "*The package is really simple to use*".

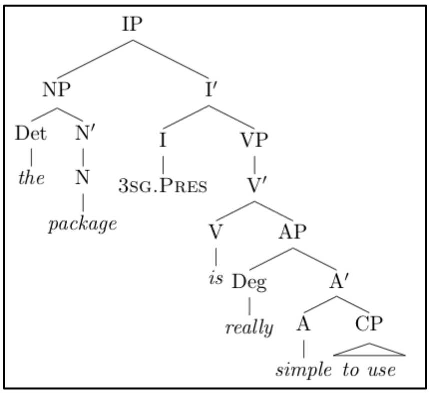


Figure 2.5 Parse Tree (Jun Li et al., 2016)

2.4.1.5 Stemming and Lemmatization

Although both stemming and lemmatization perform the same fundamental function, stemming generates 'stem' while lemmatization remits 'lemma'. Lemmatization works by capturing the valid word of the language from a dictionary without compromising on the context of the word in the sentence (Jabbar, Iqbal, Khan, & Hussain, 2018). Stemming on the other hand, returns a common variant of the word disregarding the context of the word, hence the stemmed word could be invalid. As an example, for the word *patriotic*, *patriotism, patriot*, a stemmer would return the word *patrio* but a lemmatizer identifies the root word as *patriot* as a meaningful dictionary word (Jabbar et al., 2018). Literature has shown cases where lemmatization is useful and able to assist in synonym search using thesaurus, unlike stemming (Balakrishnan & Lloyd-Yemoh, 2014).

2.4.2 Opinion Mining Phase

The opinion mining phase can be divided into three; feature extraction, sentiment orientation and opinion summarization. Each sub-phase will be discussed in detail in the following sub-sections.

2.4.2.1 Feature Extraction

Feature extraction refers to the technique of identifying distinct features or attributes in order to classify text (Varghese & Jayasree, 2013). This research is considered a form of feature-based opinion mining due to the classification process of grouping features extracted from the text into different categories. Features identified in this research refers to the purpose or intention of the extracted post being categorized into groups related to diabetes management.

2.4.2.2 Sentiment Orientation

The term sentiment orientation has been used interchangeably with opinion orientation, semantic classification and polarity in various literature (Bravo-Marquez et al., 2016; Liao, Feng, Yang, & Huang, 2016; Parkhe & Biswas, 2014). According to Fersini, Messina, and Pozzi (2016), sentiment orientation refers to the identification and categorization of opinions into positive, negative or neutral where neutral refers to a text having no opinion. Sentiment orientation can be divided into three sub-phases, namely subjectivity analysis, semantic polarity classification and polarity strength identification (Ravi & Ravi, 2015).

A. Subjectivity Analysis

L. Zhang and Liu (2017) described subjectivity as a linguistic expression of an opinion, sentiment and emotion, and the analysis of subjectivity within a written text is to detect if a given document reveals an opinion or not. The main aim of subjectivity is to separate subjective sentences from objective sentences. Subjective sentences represent opinions, evaluations and emotions etc. whereas objective sentences revolve around facts (A. Bagheri & Saraee, 2014). Take the following sentences for example:

Sentence 1: *I just got my flu shot from the doctor today.*

Sentence 2: *The flu shot is effective*

As it can be seen, Sentence 1 expresses a factual information, hence it is considered an objective sentence while Sentence 2 states an opinion of the said *flu shot*, thus it is labelled as a subjective sentence. Jijkoun, de Rijke, and Weerkamp (2010) categorized subjective sentences into on-topic and off-topic, with on-topic referring to sentences that showcase positive or negative semantic towards a specific topic.

B. Semantic Classification

Semantic classification assigns sentences/documents/features into positive, negative and neutral categories, and these categorizations can be within a numerical scale with respect to the opinions expressed by the opinion holders (Parkhe & Biswas, 2014). Studies have discovered words that encode desirable state such as beautiful, amazing, sumptuous etc. have a positive orientation, while negative orientation is assigned to words that encapsulate undesirable states such as upsetting, horrific, weak etc. This discovery has opened channels for semantic orientation based lexicons (Bravo-Marquez et al., 2016) to be generated and freely used by others working on semantic classifications.

As explained in the paragraph above, semantic classification can take the shape of a numerical range, hence semantic polarity classification can be divided into two forms: binary (bi-polar) and multi-class (fine-grained) (Habernal et al., 2014; Jotheeswaran & Koteeswaran, 2016). Semantics of sentences in a binary classification consist of three classes (i.e. positive, negative and neutral), whereas classification for multi-class is within a range (i.e. 1-n). According to L. Zhang and Liu (2017), if a semantic takes on numerical or ordinal scale values within a given range, it is then referred to as a regression problem, and if it only takes on categorical values (positive, negative and neutral) then it is considered a classification problem.

C. Polarity Strength Identification

Recent studies have looked into identifying strength of the opinion word used to express a positive or negative opinion (Barnaghi, Ghaffari, & Breslin, 2016; Devika, Sunitha, & Ganesh, 2016; Isguder-Sahin, Zafer, & Adah, 2014; Lima, De Castro, & Corchado, 2015). This is done in reference to the form of word used to identify an opinion, for example, the word *gorgeous* is more positive compared to *pretty*.

2.4.2.3 Opinion Summarization

Opinion summarization is equilateral to feature extraction and semantic classification where the purpose of this phase is to produce a synopsis of sentiment revealed in the previous stages of opinion mining (S. Sun et al., 2017).

2.4.3 Post-Processing Phase

The post-processing phase is all about data presentation in various forms and formats for the ease of users to interpret the results (Liau & Tan, 2014; Natarajan, Sankaran, Santhi, & Brindha, 2013). In some cases, data are visualized with respect to features (Chifu, Leţia, & Chifu, 2015; Lek & Poo, 2013), products (Aravindan & Ekbal, 2014; Genc-Nayebi & Abran, 2017) or even customer reviews (Claypo & Jaiyen, 2014; Srisuan & Hanskunatai, 2014). High-level data visualization of opinion mining systems assists users to make comparisons between features/products at a single glance, hence allowing producers and consumers to make informed business decisions or purchasing a product (Breen, 2012; G. Li, 2017).

The proposed classification framework of this research consists of four tiers namely sentiment, type, emotion and purpose. Therefore, in the following sections, literature studies related to each individual tier will be discussed in detail. The first section will look into sentiment analysis studies, followed by type studies, emotion and finally purpose related literature studies.

2.5 Sentiment Analysis Studies

The expedited evolution of social interaction platforms such as forums, blogs, social media, customer survey feedbacks have generated a colossal amount of expressive data for analysis. Such data can be beneficial to numerous domains such as business, entertainment, political, health, governments etc. An overview of sentiment analysis and its approaches have already been discussed in Chapter 1; therefore, this section will delve into reviewing previous researches in the field of sentiment analysis.

The trend of research within this field has diversified in many aspects especially with the discovery of new methods of analyzing different forms of text, audio and visual formats (D. Jiang, Tao, Wang, & Dong, 2019; Poria et al., 2017; Poria, Cambria, Howard, Huang, & Hussain, 2016). However, this research is interested in the different approaches applied to analyze textual data, hence the focus of this section will be on studies that have used textual corpus. Table 2.1 shows some of the recent studies in sentiment classification

Table 2.1 Sentiment Analysis Studies

Reference	Classification Type	Technique	Data Set	Corpus Language	Results
Al-Smadi, Al-Ayyoub, Jararweh, and Qawasmeh	Aspect	Naïve Bayes, Decision Tree, K-Nearest Neighbor, Support Vector Machine	SemEval-2016: Task-5	Arabic	Best accuracy achieved by Support Vector Machine = 59% for target extraction

Jiménez-Zafra, Martín-Valdivia,	Document	Support Vector Machine, Lexicon Dictionary	Drug and Physician Reviews	Spanish	Drug reviews more difficult to classify than physicians
Elhadad, Li, and Gebali (2019)	Document	Vector Space Model, Decision trees, Naive-Bayes, Logistic Regression, Perceptron, Multilayer Perceptron	Twitter	Arabic, English	Best accuracy = 84% for Naïve Bayes for Uber rides related tweets
García-Pablos, Cuadros, and Rigau	Aspect	W2VLDA for topic modelling	SemEval 2016 task 5 (ABSA)	English, Dutch, Spanish	Improved classification for aspects and polarity classification
Pham and Le (2018)	Aspect	Compositional vector model, Back-propagation algorithm based on Gradient Descent	Hotel Reviews	English	Proposed model outperforms other popular methods
T. Chen, Xu, He, and Wang (2017)	Sentence	BiLSTM-CRF	Stanford Sentiment Treebank	English	Proposed model produces better classification accuracy
Appel, Chiclana, Carter, and Fujita (2016)	Sentence	Proposed Hybrid Method	Movie Reviews	English	Hybrid approach produce higher accuracy compared to Naïve Bayes
Tripathy, Agrawal, and Rath (2016)	Document	Naive Bayes, Maximum Entropy, Stochastic Gradient Descent, Support Vector Machine	Movie reviews and blogs	English	As the value of 'n' in n-gram increases the classification accuracy decreases
Le, Van Le, and Pham (2015)	Aspect	GK-LDA for topic modelling	Product reviews	Vietnamese	proposed method effectively performs aspect and classification task

It can be observed from the table above that most popular approach remains to be the machine learning approach. However, the classification form differs from one another and so does the corpus language. The present research drew inspiration from previous studies and adopted the machine learning approach in order to analyze the sentiment of the posts extracted. However, a lexicon dictionary, SentiWordNet 3.0 (Baccianella, Esuli, & Sebastiani, 2010), was also used to classify sentiments. SentiWordNet 3.0 was selected instead of other available dictionaries such as SentiStrength (Thelwall, 2013) or Linguistic Inquiry and WordCount (LIWC) (Pennebaker, Francis, & Booth, 2001) as it is extensive and is able to analyze sentiments regardless of word length of a given post.

2.6 Type Classification Studies

Type classification in this research refers to classifying the extracted posts into one of three known types of diabetes, namely Type 1, Type 2 and Type 3 (i.e. gestational diabetes). Past studies on diabetes have been more medical prone with scientific description and diagnosis on how to distinguish one type from another (Association, 2005, 2010). El-Sappagh and Ali (2016) conducted a study on the ontology aspects of classifying diabetes, however the ontology was only prepared for Type 1. Studies in classifying diabetes according to its distinct types, and analyzing each type individually from a sentiment, emotion and purpose perspective has yet to be conducted. Therefore, the present research aims to fill these gaps by classifying extracted diabetes data based on sentiment, emotion and purpose, for each types of diabetes.

2.7 Emotion Analysis Studies

Similar to sentiment analysis, the evolution of social networks has given rise to means of how people interact and express their emotions through these channels. Compared to audio and video components, literature has shown that text is still the most common form of communication on social media platforms (Sailunaz, Dhaliwal, Rokne, & Alhajj, 2018). Detecting emotion from a given text would be fairly easy if words representing the said emotion were explicitly mentioned. However, in most cases, the emotion expressed is done in a subtle form, and a sentence may carry more than one form of emotion within it. For example:

"I can't believe how I ended up with diabetes with this pregnancy, feels like I'm being punished"

The emotion observed in the sample sentence above is *surprise* as the author cannot believe he/she has been diagnosed with diabetes and *anger* as he/she feels punished. Over the years, a considerable amount of effort was taken to produce automated emotion detection systems to correctly identify and classify human emotions from text. Table 2.2 provides some of the recent studies within this domain.

Table 2.2 Emotion Analysis Studies

Reference	Technique	Emotion Detected	Results
Chatterjee, Narahari, Joshi, and Agrawal	Compared different Neural network models using benchmark text	Happy, Sad, Angry	Most systems tested had best performance for *Sad* emotion class, and worst for *Happy* emotion class
Hasan, Rundensteiner, and Agu	Develop a supervised learning system called Emotex	Happy, Unhappy	Proposed model was able to correctly classify 90% of emotions
Rout et al. (2018)	Multinomial Naïve Bayes with lexicon dictionary	Anger, Fear, Joy, Love. Sad, Surprise, Thankfulness	80.68% accuracy achieved using proposed method
An, Sun, and Wang (2017)	Naïve Bayes	Comfortable, Happy, Inspirational, Joyful, Lonely, Miss, Nostalgic, Passionate, Quiet, Relaxed,	Result for English lyrics not as accurate as Chinese lyrics

		Romantic, Sad, Soulful, Sweet, Yearning	
S. Mohammad and Bravo-Marquez	Regression	Anger, Fear, Joy, Sadness	Accuracy not verified
Perikos and Hatzilygeroudis (2016)	Ensemble of Naive Bayes, Maximum Entropy and Knowledge-based tool	Circumplex model	Lack of lexical resources, accuracy is not better than all baseline methods
Bu, Li, Cao, Wu, and Zhang	Proposed Emotional Evolution Prediction Algorithm	Happy, Popularity	Average precision = 85% compared to baseline
W. Li and Xu (2014)	Support Vector Machine and Support Vector Regression	Happiness, Anger, Disgust, Fear, Sadness, Surprise	Best results (85%) achieved for Happiness classification
Desmet and Hoste (2013)	Support Vector Machine with bootstrapping	Abuse, Anger, Blame, Fear, Forgiveness, Guilt, Happiness, Hopefulness, Hopelessness, Love, Pride, Sorrow, Thankfulness	Was able to detect frequent emotions but rare emotions results were poor

From the table (Table 2.2), most of the techniques adopted in the past have combined machine learning and lexicon-based approaches. However, this yielded in low accuracy results especially in identifying multiple emotions as past studies looked to identify a complex amount of emotion which lacked training data for each emotion. The present research aims to improve emotion detection techniques within the multi-tier framework in order to achieve a better accuracy for multiple emotions.

2.8 Purpose Analysis Studies

S. M. Mohammad et al. (2015) defined purpose as the hidden human intention behind a tweet or post. A study related to airline industry revealed customers used social media platforms to complaint regarding their ticketing issues or delay in flight time and also to make suggestions on improving the on ground customer service facilities available (W. Kaur & Balakrishnan, 2018). In another political related domain study, the purpose found within the text extracted revealed the purpose of users to post on social media platforms was to instigate, to vent and also to show support for the candidate at hand (S. M. Mohammad et al., 2015).

Using the same definition, this research aims to classify posts extracted according to the reason(s) behind the post. In the context of diabetes for example, a purpose can be to seek information regarding treatment or therapy, to share a diabetic friendly recipe, or to seek help regarding a diagnosis etc. Table 2.3 shows a sample of purpose classifications based on the corpus used for this research.

Table 2.3 Sample Purpose Classification Text

Sample Post	Purpose
Unfortunately, yes, I occasionally need prednisone when I have the common cold because respiratory viruses cause my asthma to flare up. My goal during that time is just to keep BG under 200 which is very difficult. Prednisone really antagonizes insulin. It is very frustrating.	To provide feedback on drug options
Using applesauce and whole-wheat flour makes these decadent muffins far healthier than your typical baked good!	To share diabetic friendly recipes
Yes, everybody and every struggle is different. We all battle the same disease just in our own ways. Don't let someone not tell you are different than someone else with T1	To provide emotional support

PCOS gets better, even disappears and child bearing readily happens on a Ketogenic/Low-carb way of eating. - - Food is medicine. Always has been. How can we expect to heal ourselves when we eat food-fill ingredients we can't even pronounce?	To vent frustration

Literature review shows that purpose or intent classification using text as corpus still has a wide scope for improvement. Table 2.4 shows a list of studies on purpose classification, with the majority having used machine learning algorithms. It is to note that only the study of Wang et al. (2019) produced a high accuracy (i.e. 91%), which was using chatbot and dialogues as an input dataset. The aforementioned research proposed a hybrid approach of two different neural networks in order to classify dialog utterance between chatbot and users in identifying their intent of talking to a chatbot. Another study by Purohit, Dong, Shalin, Thirunarayan, and Sheth (2015) looked to identify purpose of tweets using AdaBoosting feature selection techniques which produced an accuracy of 68%. Therefore, the present research intends to examine other techniques to improve classification accuracy for purpose.

Table 2.4 Purpose Classification Studies

Reference	Research Objective	Technique	Results
Y. Wang, Huang, He, and Tu (2019)	To propose a hybrid architecture to classify the intent of a dialogue utterance	Convolutional neural network and bidirectional gated recurrent unit neural network (CNN-BGRU)	Precision achieved is 91%
Setyawan, Awangga, and Efendi (2018)	To identify intent on the chatbot system	Logistic Regression, Multinomial Naïve Bayes	Logistic regression achieved better accuracy (72%) compared to Multinomial Naïve Bayes
Agarwal and Sureka (2016)	To develop a cascaded ensemble learning classifier for identifying posts with	Random Forest, Naive Bayes, Decision Tree	Data for intent classification was imbalanced hence accuracy obtained was less than 12%

	racist or radicalized intent		
Tsur, Pinter, Szpektor, and Carmel (2016)	To detect queries with a question intent	Random Forest	The results were inconclusive
S. M. Mohammad et al. (2015)	To determine sentiment, emotion and purpose of tweets	Used human annotation to determine purpose for training data. Used Support Vector Machine for automatic classification	Precision achieved for purpose classification = 58.3%
Purohit et al. (2015)	To classify intent using social media data during crisis event	Used bottom up and top down processing with knowledge guided patterns	The best accuracy achieved was 68%

The dataset extracted for this research comes from online resources related to diabetes. Hence, the following section will look to provide a brief introduction on prevalence of diabetes both globally and within the context of Malaysia as well as a review on literature related to online health communities and Facebook.

2.9 Online Health Communities - Facebook

Diabetes is among the largest global health emergencies that have long been overlooked. It is a persistent condition that arises when the body is either unable to provide enough insulin or unable to use insulin. This disease is diagnosed by monitoring blood glucose levels and the ineffectiveness of the body to regulate insulin means glucose remains circulating within the bloodstream. Continuous high levels of glucose in the blood (also known as hyperglycemia) not only causes harm to tissues in the body but may also lead to life-threatening complications (Group, 1979).

According to the 2017 global report produced by World Health Organization, more than half a million of world population of children under the age of 14 are reported to be battling Type 1 diabetes. Apart from this, 415 million adults are already being treated for diabetes with another 318 million traced to have glucose intolerance, which ultimately put them at a high risk of contracting diabetes in the future. These numbers are predicted to exceed 642 million people by 2040. Bringing these numbers closer to home, the National Health and Morbidity survey of 2015 conducted by the Ministry of Health Malaysia have discovered one in every five Malaysians to be diabetic, with the prevalence showing an alarming upward trajectory (Figure 2.6).

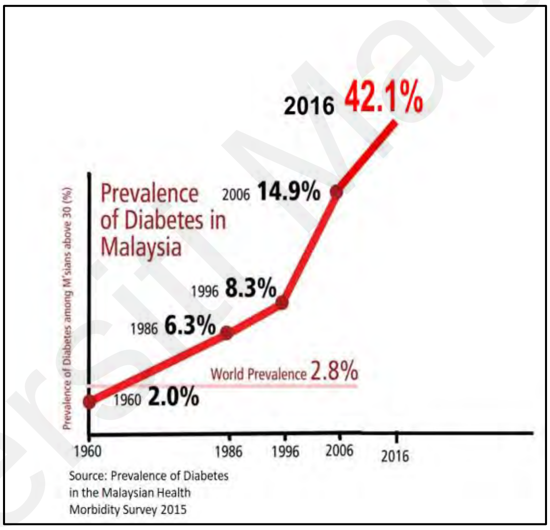


Figure 2.6 Prevalence of Diabetes in Malaysia (National Health and Morbidity Survey, 2015)

According to Abedin et al. (2017), it is not just diabetes patients who need to have periodic contact with health professionals, in some cases, even caregivers are in contact with doctors and clinicians (e.g. parents). As such, patients and caregivers are in need of both skills and support to manage the disease.

The growing expenses of medical treatment and the urge to seek alternative medication options has prompted users to turn to online health communities as an avenue to reach out to patients and caregivers (McRoy et al., 2018). Online communities have evolved from chat rooms to listservs, bulletin boards, news groups and web forums to social media platforms since its inception in the 1990s. Members of the community who are battling the same health conditions such as cancer (Gage-Bouchard, LaValley, Mollica, & Beaupin, 2017; McRoy et al., 2018), mental health conditions (Fergie, Hilton, & Hunt, 2016; Maestre, Herring, Min, Connelly, & Shih, 2018) and other chronic diseases (Willis & Royne, 2017) etc. are able to seek others for support via social media. Studies have suggested such communities serve as a priceless warehouse of information exchange, and personal stories ranging from discussions on symptoms and diagnoses to medications and side effects. Communications sometimes also include suggestions on doctors, clinics, financial assistance and daily life advice (Greene, Choudhry, Kilabuk, & Shrank, 2011; Y. Zhang et al., 2013).

One such platform is Facebook. According to the Global Web Index Flagship 2017[8] report, Facebook dominates the social landscape with the greatest number of traffic recorded compared to YouTube, Instagram and Twitter. It has also been reported to be popular among all income and education level users. On average, 70% of the world population has reported to log on to Facebook on a daily basis with over 43% doing so several times a day on a daily basis. This is also supported by another internet survey conducted by the Malaysian Communication and Multimedia Commission[9] that has identified Facebook as the most used social networking site in 2017 with over 97.3% Malaysian population using this site for information exchange and sharing personal

[8]http://www.upa.it/static/upload/soc/social-summary-report-new.pdf
[9]https://www.mcmc.gov.my/skmmgovmy/media/General/pdf/MCMC-Internet-Users-Survey-2017_v2.pdf

messages. This acumen is used as the basis of using Facebook as a source of data extraction for the purpose of this research.

The following sub sections will look into studies conducted using Facebook as a data source for sentiment and emotion analysis studies.

2.9.1 Facebook Sentiment and Emotion Analysis

A vast majority of the studies on sentiment and emotion classification have focused on the aspect classification of either sentiment, or emotion (Akter & Aziz, 2016; Alashri et al., 2018; Krebs, Lubascher, Moers, Schaap, & Spanakis, 2017). Studies addressing more than one aspect are few, with the exception including the work of S. M. Mohammad et al. (2015) who classified electoral related tweets based on their sentiment, emotion, purpose and style. The present research differs from S. M. Mohammad et al. (2015) as the pre-processing of tweets differs compared to Facebook posts (Salloum, Al-Emran, Monem, & Shaalan, 2017). Furthermore, S. M. Mohammad et al. (2015) analyzed the tweets for sentiment, emotion, purpose and style aspects separately , whereas this research combines four elements (sentiment, type, emotion and purpose) within a cascading multi-tier framework to achieve a higher percentage of accuracy. Some studies within this domain have also analyzed sentiments with respect to different languages such as Arabic (Akaichi, Dhouioui, & Lopez-Huertas Perez, 2013), Vietnamese (Trinh, Nguyen, Vo, & Do, 2016), Bengali (Islam, Islam, Hossain, & Dey, 2016) and Portuguese (Rodrigues et al., 2016) etc.

Poria, Cambria, and Gelbukh (2016) presented a seven-layer deep convolutional neural network model in one of the earliest approaches to aspect extraction using deep leaning approach. The idea was to detect particular aspects of a service or product that the opinion holder was specifically praising or complaining about by tagging each word of the review as either an aspect or non-aspect word. The proposed classifier used word embeddings of linguistic patterns and trained the data using neural networks to achieve higher accuracy compared to other state-of-the-art approaches. X. Sun et al. (2018) on the other hand adopted multivariate Gauss distribution to analyze user emotion to detect abnormal emotional state from social media text for four emotions namely happy, surprised, sad and angry. The accuracy achieved by proposed model was 83.49% for the emotion happy. Table 2.5 and 2.6 showcase studies that have analyzed Facebook posts for sentiment and emotion analyses, respectively.

Table 2.5 Facebook Sentiment Analysis Studies

Reference	Classifier	Facebook posts language	Model Adopted	Results
Dhaoui, Webster, and Tan (2017)	Maximum Entropy, Random Forest, Support Vector Machine, Bagging and Decision Tree	English	LIWC2015 lexicon and RTextTools	Combining both approached produced accuracy of 66.3%
Rodrigues et al. (2016)	AlchemyAPI, Semantria, Sentistrength and TextAnalytics	Portuguese	Proposed SHC-pt model	Proposed tool provided the best accuracy of 58%
Trinh et al. (2016)	Support Vector Machine	Vietnamese	Own model with Vietnamese lexicon dictionary	Highest accuracy achieved for sport category (95%)
Ortigosa et al. (2014)	J48, Naïve Bayes, Support Vector Machine	Not mentioned	Proposed SentBuk	Proposed model achieved accuracy of 83.27%
Troussas, Virvou,	Naïve Bayes, Rocchio and Perceptron	English	Own model	Naïve Bayes produced best accuracy of 77%

With respect to the table above, there have been studies in the past that have looked into capturing Facebook behavioral data (likes, comment, shares and reaction) to analyze for sentiment and emotion. Nevertheless, the studies that are related to sentiment were mostly content analysis papers that looked into the impact of using such behaviors in determining the overall communication patterns within the organization or how the usage of such behaviors indirectly provided publicity for branding thus increasing sales (C. Kim & Yang, 2017; Quesenberry & Coolsen, 2018). Calero (2013) found Facebook gives different weights to different behaviors to regulate what appears on a user screen. A share is given the highest weight, as much as two comments and seven likes. This finding implies Facebook acknowledges how each behavior differs from each other.

Table 2.6 Facebook Emotion Analysis Studies

Ref	Methodology	Model Adopted	Results
Corazza et al. (2018)	Used EMOLEX to determine words linked to emotions	Recurrent neural network	Model was able to detect hate speech with 67.5% accuracy
Sandoval-Almazan and Valle-Cruz (2018)	Facebook reactions were indexed as sentiment related to gauge voter's emotion	Sentiment Index	Voters sentiment on Facebook was not good indication for political outcome
Krebs et al. (2017)	Used Facebook reactions to map emotions	Convolutional neural network and Recurrent neural network	Use of reactions were able to gauge customer emotions more accurately
Bazarova, Choi, Schwanda Sosik,	Used posts to investigate how emotion is shared on Facebook	Statistical analysis	People share less positive emotional content over network visible communication
Sherawat (2015)	Proposed EmoMint algorithm to detect anger and fear emotion from posts	Own model	The proposed model was able to detect anger and fear correctly for longer posts

The table above shows past studies related to emotion mostly related to detecting emotions within the corpus instead of classifying the strongest emotion. Therefore, literature has mostly revolved around improving emotion detection instead of identifying predominant emotion within corpus and classifying it. The following sections explains each behavior individually,and discusses the related studies.

2.9.2 Facebook Feature: Like

Ekström and Östman (2015) uncovered two modes of communication that occur on Facebook, namely active interaction (liking, sharing, commenting and reacting) and passive interaction (clicking, watching, viewing/hovering). This research is interested in using the number of *like*, *share*, *comment* and *reaction* as such nuances are publicly available and can be used to arbitrate sentiment and emotion strength. Additionally, recent literature indicates such features can be useful in measuring human interaction on social media (Carah, 2014; Ding, Cheng, Duan, & Jin, 2017; Ross et al., 2018).

The like button available on Facebook has been a part of Facebook since its birth; enabling users to quickly interact with status updates, comments, links etc. A click of the like button immediately distributes the said content within one's newsfeed while displaying the number of other users who have followed suit to like the same content (Ding et al., 2017; Sumner, Ruge-Jones, & Alcorn, 2017). Sumner et al. (2017) found the like button was not only used to imply agreement with the content shared or express their preference to their friends, but it was also used as a tool to provide positive feedback, particularly in liking advertisements. For instance, users may hit the like button for a post, prompting their friends to do the same, and thus strengthening their views on the post.

In a marketing study conducted by Pelletier and Blakeney Horky (2015), it was found that users were eight times more prone to click the like button collated to share or comment. There was a 44% chance of an item to be liked at least once a day, with 29% of Facebook users actively liking contents throughout the day. Recent literature has shown the impact of like in indicating preference and how one feels about a content that is to be shared (Coursaris et al., 2016; Vepsäläinen et al., 2017; Zell & Moeller, 2018). This is also reiterated in Sumner et al. (2017), implying positive feedback when the like button is hit, ergo, the more the number of likes a content receives, the stronger (more positive) the sentiment it perceives. Hence, the present research hypothesizes that *like* improves sentiment intensity.

2.9.3 Facebook Feature: Share

Despite the ease to share information across Facebook via the like button, Facebook EdgeRank[10] Algorithm has placed a higher significance on *share,* hence indicating a greater commitment in sharing content on Facebook compared to just liking it (C. Kim & Yang, 2017). A particular content on Facebook can be shared either by reposting it on one's own timeline, a friend's timeline or sending it as a private message. Regardless how a content is shared, each time a user clicks on the share button, the number is recorded and displayed on the content itself.

Studies have found a substantial link between the type of content a user shares with the users self-image presentation (Cvijikj & Michahelles, 2013; J. R. Rui & Stefanone, 2013). P.-W. Fu et al. (2017) found users are more inclined to share a content that is an extension of what they feel and think and how they would like themselves to be perceived by their friends. Hence, users are more aware of the content they are sharing, and thus indicating a higher cognitive evaluation compared to liking. Moreover, the shared content

[10]http://edgerank.net/#How-does-EdgeRank-work

is not only visible to their own circle of friends but also to a greater reach of friends on Facebook, thus allowing more visibility (Carah, 2014; P.-W. Fu et al., 2017; C. Kim & Yang, 2017).

Taking into account the conscious awareness that goes into sharing a particular content on Facebook (P.-W. Fu et al., 2017), it goes to show the importance of the share button in determining a sentiment strength, as people are more inclined to share when they are truly in agreement of the content and consciously support it.

2.9.4 Facebook Feature: Comment

Facebook comments are user generated content published as a thread under a single item posted by users. Commonly, comments can be viewed by all users who are connected within the same network (friends of users), however, comments can be kept private by adjusting settings within one's user profile. Users are allowed to *like* and *share* comments just as they can with Facebook posts.

Most sentiment analysis studies related to comments have analyzed each comment individually as they would with posts (Akaichi, 2014; Ortigosa et al., 2014; Trinh et al., 2016). Comments play an imperative role specifically in impacting readers' understanding on crucial discussions (Hong & Cameron, 2018) or influencing voters' opinions during electoral polls (Alashri et al., 2018; H.-C. Lin, 2017). In a brand awareness and engagement research, Carah (2014) found comments have a tendency to sway public opinion as readers are able to browse through the comment thread of a discussion, which eventually influence their purchase intention. The present research assumes that a high number of comments within a thread indicates a heated sentiment (positive or negative). Moreover, comments will be analyzed individually for sentiment and emotion, by taking the frequency of comments into consideration as well.

2.9.5 Facebook Feature: Reaction

Facebook rolled out the reaction buttons in February 2016 as an extension of its like button. The reaction buttons are represented as emojis (Figure 2.7), each representing a graphical expression of one's emotion (Ye Tian, Galery, Dulcinati, Molimpakis, & Sun, 2017). Reactions work in the very same way as the like button and the options (love, haha, wow, sad and angry) appear when one hovers on the like button.

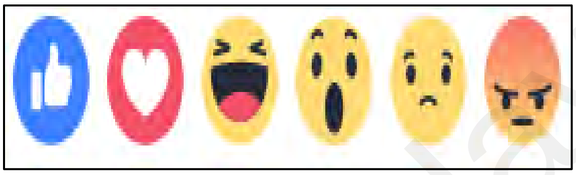


Figure 2.7 Facebook reaction emojis

Pool and Nissim (2016) believes reaction buttons allow users the luxury of reacting to a post, comment or share by expressing themselves wordlessly. This permits users to communicate their emotions without having to spell it out. A study by Smieško (2016) found that the reaction button was accepted as a substitute of emotion expression in the digital world, and it has been accepted as a form of modern speech by the United States Court of Appeals. Furthermore, literature in recent times have begun to explore the use of reactions in emotion analysis. For example, Krebs et al. (2017) predicted the distribution of reaction for a new post and how this affects customers emotions. Neural network algorithms were used, with results indicating reaction distribution prediction with a mean squared error of 0.135. Ross et al. (2018) on the other hand, studied the form of reaction posted on Facebook after the Berlin, London and Stockholm terrorist attacks, with results showing the page administrators disseminating news and information for crisis situation were able to use the number of reactions to predict the form of information that should be uploaded. Considering the impact of Facebook reactions on these studies, the present research hypothesizes that sentiment intensity is enhanced when reactions are considered.

Table 2.7 shows studies that have used Facebook features (like, share, comment and reaction) for analysis. Although numerous studies have been conducted taking these Facebook features into consideration, yet to the best of our knowledge, studies converting these features to determine sentiment intensity has not been conducted. Therefore, this research aims to extract each of these features to determine sentiment and emotion intensities for an improved classification.

Table 2.7 Facebook Features Used for Analysis

Ref	Objective	Facebook Behavior	Technique	Results
Quesenberry and Coolsen	To investigate how number of like, comment and shares impacts marketing campaign	Like, Comment, Share	Content Analysis	Increased number of shares, comments and likes is an integral part to marketing campaigns on Facebook
Zell and Moeller (2018)	To investigate importance of *like* and *comment* towards sentiment	Like, Comment	Content Analysis	Receiving more *like* on a post was associated with stronger sentiment
Ding et al. (2017)	Studied the impact of pre-release *like* to the number of sales in box office	Like	Content Analysis	The higher the number of *like* on Facebook the better opening was recorded during opening week.
C. Kim and Yang (2017)	To investigate how like, comment and share differ in behavior and usage	Like, Comment, Share	Statistical analysis	*Like* is affectively, *comment* is cognitively triggered, and share is either affective or cognitive or a combination of both.
Ye Tian et al.	To investigate the use of reactions to indicate sentiment and emotion	Reactions	Lexicon and Naïve Bayes	Reactions and emojis can be used to detect sentiment and emotion
Vepsäläinen et al. (2017)	To analyze if number of *like* could be used to predict the outcome of the parliamentary election	Like	Statistical analysis	The number of *like* and votes were shown as a positive relationship

Pool and Nissim (2016)	Use reaction to train algorithm for emotion detection	Reaction	Support Vector Machine, Naïve Bayes	Accuracy is higher for NB (85.5%) compared to SVM (82.3%)
Turnbull and Jenkins (2016)	Use reaction to engage consumers emotional connection to campaigns	Reaction	Content Analysis	Use of reactions help marketers understand how consumers emotionally connect with content displayed
Carah (2014)	To investigate how like, comment and share can impact brand activity on Facebook	Like, comment, share	Content Analysis	The more users engaged with a brand using those features, the brand made more sales

2.10 Naïve Bayes

This research adopts the Multinomial Naïve Bayes classification algorithm, which is a semi-supervised machine learning algorithm. Hence, a brief literature introduction into the classifier is provided in this section.

The Naïve Bayes classifier is a probabilistic classifier based on Bayes theorem with naïve independence assumptions. According to literature, it is one of the most basic classification techniques available with applications throughout multiple domains such as email spam detection, document classification, explicit language detection as well as sentiment classification (Medhat et al., 2014). In spite of its naïve design and simplistic assumptions, the application of this algorithm has contributed in overcoming many complex real-world classification problems such as tracing the usage of marijuana in treating post-traumatic stress disorder, detecting racial slurs within social media text and so forth (Dai & Hao, 2017; Duwairi & Qarqaz, 2014; Mukherjee & Bala, 2017; Rohani & Shayaa, 2015). The Naïve Bayes classifier may often be outperformed by other more advanced algorithms such as boosted trees, Random Forest, Support Vector Machine etc., however, it is more efficient in terms of computation time (both CPU and memory), and it requires only a small amount of training data (Diab & El Hindi, 2017).

Literature shows several variations to the Naïve Bayes classifier such as Multinomial Naïve Bayes, Binarized Multinomial Naïve Bayes, Bernoulli Naïve Bayes and Gaussian Naïve Bayes etc. (L. Jiang, Zhang, Li, & Wu, 2018; Verma & Thakur, 2018).Multinomial Naïve Bayes classifier has shown promising results in classifying words with respect to intensity including capitalization sensitiveness and multiple occurrences of a word (Diab & El Hindi, 2017). The Binarized Multinomial Naïve Bayes on the other hand, does not take the number of occurrences into consideration but rather focusses on the fact that a particular word has been used.

The following subsections looks into a brief theoretical background on Naïve Bayes algorithm as well as the adopted Multinomial Naïve Bayes.

2.10.1 Theoretical Background of Naïve Bayes

According to Harry Zhang (2004), the Naïve Bayes algorithm is based on a "naïve" assumption of independence between every pair of features. It assumes all features are independent of each other within the context of the class. It is due to this assumption, that the Naïve Bayes algorithm avoids structural learning which leads to a more simplified parametrical learning specifically when the number of features identified are very large (L. Jiang, Wang, Li, & Zhang, 2016). Although the algorithm has a tendency of over estimating the probability of a selected class, the decision making factor of the classifier is correct thus proving the accuracy of the model (Harry Zhang, 2004). For example, given a class variance y and a dependent feature x_1 through x_n, Bayes theorem states the following relationship:

$$P(y|x_1, \ldots., x_n) = \frac{P(y)|P(x_1, \ldots., x_n|y)}{P(x_1, \ldots x_n)} \quad \text{Eq. (1)}$$

Using the naïve independence assumption that

$$P\big(x_i\big|y, x_1, \ldots., x_{i-1,}x_{i+1,} \ldots, x_n\big) = P(x_i|y)$$

For all i, this relationship is simplified to

$$P(y|x_1, \ldots., x_n) = \frac{\mathrm{P}(y)\Pi_{\mathrm{i=1}}^{\mathrm{n}}\mathrm{P}(x_i|y)}{P(x_1, \ldots x_n)} \quad \text{Eq. (2)}$$

Friedman, Geiger, and Goldszmidt (1997) states a Bayesian model comprises of both a structural model as well as conditional dependencies with random variables. Therefore, a feature extracted for classification is also included within the text upon which a classification is made. For example, the sentence "I love this brand", for which the features are *I*, *love*, *this*, *brand*, where all the stated features are assumed to be conditionally independent in Naïve Bayes with respect to the class label. This assumption then allows the classifier to be defined as:

$$P(y|x_1, \ldots., x_n)\ \alpha\ P(y) \prod_{i=1}^{n} P\ (x_i|y)$$

$$\hat{y} = \arg max\ \ P(y) \prod_{i=1}^{n} P(x_i|y) \quad \text{Eq. (3)}$$

2.10.2 Multinomial Naïve Bayes

The Multinomial Naïve Bayes (MNB) is among the typical Naïve Bayes alternative used for classification, specifically for data that are depicted as word vectors (Frank & Bouckaert, 2006). The distribution is defined by vectors $\theta_y = (\theta_{y1}, \ldots., \theta_{yn})$ for each class *y*, where *n* is the number of features (size of vocabulary) and θ_{yi} is the probability

$P(x_i|y)$ of feature i appearing in sample of class y. Frank and Bouckaert (2006) found parameters θ_y can be further estimated by a smooth version of maximum likelihood as shown in the equation below:

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \qquad \text{Eq. (4)}$$

Where $Ny_i = \sum_{x \epsilon T} x_i$ refers to the number of times i appears in sample class y in training set T. Therefore, the total count of features for class y can be defined as:

$$N_y = \sum_{i=1}^{|T|} N_{yi} \qquad \text{Eq. (5)}$$

Text classification adopting MNB has been studied in length in the past especially in classifying multi-label text. Multi- label text refers to those documents that can be classified into two different documents. For example, in the case of diabetes, a document classified as 'Treatment' can be labelled as 'Dietary changes' as well. Although MNB has been regarded as an efficient and reliable classifier specifically for multi-label classifications, the performance of it is still lacking in comparison with Support Vector Machine or other boosting algorithms (C.-H. Lee, 2018). However, literature has shown substantial improvement when MNB is integrated with feature selection methods of weights (C.-H. Lee, 2018; Wongso, Luwinda, Trisnajaya, & Rusli, 2017). However, there is still space for improvement especially in improving class dependencies by exploiting vector types from binary (numeric) to string. This will be further explained in the following subsection. Table 2.8 displays some of the latest literature using Multinomial Naïve Bayes model.

Table 2.8 Studies adopting MNB model for text classification

Reference	Technique	Type of Vectors	Results
Dhar, Mukherjee, Dash, and Roy (2018)	Multi-Layer Perceptron, Random Forest, Support Vector Machine, MNB and KStar	Unknown	Highest accuracy achieved by Multilayer Perceptron while MNB came in fourth
C.-H. Lee (2018)	Used MNB with proposed value weighting method for feature selection	Numeric	Accuracy of classification improved by 12%
Setyawan et al. (2018)	MNB, Logistic Regression	Numeric	Logistic regression achieved better accuracy (72%) compared to MNB
Wongso et al. (2017)	Feature selection – Singular Value Decomposition and TF-IDF MNB, Multivariate Bernoulli and Support Vector Machine	Numeric	TF-IDF and MNB gives the highest accuracy (85%) compared to rest.
L. Jiang, Li, Wang, and Zhang (2016)	Propose algorithm called multinomial naive Bayes tree (MNBTree) by deploying a multinomial naive Bayes text classifier on each leaf node of the decision tree	Numeric	Proposed algorithm achieved better accuracy compared to MNB alone
Zhou, Tong, Gu, and Gall (2014)	MNB combining text mining and data mining techniques for classification	Numeric	Accuracy = 56%
Ibrahim and Landa-Silva (2014)	Used MNB with frequency transforming feature selection	Numeric	Accuracy = 58%

2.11 Numeric and String Vectors

As it can be observed from Table 2.5, classification studies in the past have opted to use numerical vectors compared to string vectors, resulting in problems such as huge dimensionality and sparse distribution (Allahyari et al., 2017; Jain & Mandowara, 2016;

Jo & Cho, 2008; H. Kim, Howland, & Park, 2005). This in return affects the performance of the classification algorithm, and eventually leading to low accuracy readings. This section looks into the concept of Vector Space Model, string vectors and the proposed classification weighted algorithm.

2.11.1 Vector Space Model

In order to gain a better understanding on the description of the problem, it is necessary to define some terms and variables used frequently within this section for clarification purposes. Given a collection of documents D = (d1, d2, …, dD), let V = (w1, w2, … wv) be the distinct words/terms in the collection. Hereon forth, V will be referred to as the vocabulary. Frequency of word w ∈ V in document d ∈ D is demonstrated by fd(w) and the number of documents containing word w is depicted by fD(w). Hence, $\vec{t}_d$= (fd(w1), fd(w2), …, fd(wv)) represents the term vector for the document. The most common representation of these documents is to convert them into numeric vectors known as Vector Space Model (VSM) (Allahyari et al., 2017).

VSM was initially introduced for indexing as well as information retrieval (Salton, Wong, & Yang, 1975), however, recent uses of VSM has included text mining and analysis of large collection of documents (Allahyari et al., 2017; Zhai, 2017). In a VSM model, each variable that has been assigned a numeric value indicates the weight (importance) of the word within the document. According to Allahyari et al. (2017), there are two main term weight models:

a) Boolean Model

In this model, a weight $w_{ij} > 0$ is assigned to each word $w_i \in d_j$. For any term that does not appear in the document d_j, the default weight assigned is zero.

b) Term frequency-inverse document frequency (TF-IDF)

TF-IDF is the most frequently used weighting schemes in text classification (Ahmad, 2017; Aravindan & Ekbal, 2014; Ghorpade & Ragha, 2012). Let q be the term weighting scheme and the weight of each word w ∈ d is computed as follows:

$$q(w) = f_d(w) * \log \frac{|D|}{f_{(D)} w} \quad \text{Eq. (6)}$$

Where |D| refers to the number of documents in the collection D

In TF-IDF, the term frequency is normalized by the inverse document frequency, IDF. This normalization is needed to contract the weight of the terms appearing more commonly in the document (Mazzonello, Gaglio, Augello, & Pilato, 2013). This is to ensure the words that appear less within the document are also taken into consideration during the classification process. With respect to the weighting scheme, each document is then represented by a vector of term weights w(d) = w(d, w1), w(d, w2), …, w(d, wv)). Hence the similarity between both documents (d1, d2) can be computed using the cosine similarity as follows (Allahyari et al., 2017):

$$S(d1, d2) = \cos(\theta) = \frac{d1 \, . \, d2}{\sqrt{\sum_{i=1}^{v} w_{1i}^2} \cdot \sqrt{\sum_{i=1}^{v} w_{2i}^2}} \quad \text{Eq. (7)}$$

Rahmawati and Khodra (2016) used TF-IDF and bag of words word2vec algorithm, and found the accuracy of newspaper article classification to have improved from 76.73% to 80.17%. Similarly, Xue, Fu, and Shaobin (2014) built a semantic dictionary for the Chinese language using cosine similarity, while D. Zhang, Xu, Su, and Xu (2015) used SVMperf algorithm with word2vec to classify social media messages on Sina Weibo.

2.11.2 String Vectors

The concept of string vectors have always been apparent within the machine learning process, however, it is not until recently studies have looked into manipulating them to reduce dimensionality issues (Allahyari et al., 2017). The most commonly used model to compute vector representation of words is known as Word2Vec (Chakraborty, Bhattacharyya, Bag, & Hassanien, 2018). There are two main learning algorithms within this model; the continuous bag of words and continuous skip gram, both of which work by learning the representation of the word that would be useful in predicting other words in a sentence.

In brief, Word2Vec is a machine learning algorithm that produces "word embeddings". The idea of Word2Vec is to represent words in a vector space (as machine learning algorithms are only be able to understand numeric terms), and to bridge the gap between each word to improve word prediction (Chakraborty et al., 2018). For example, assume vector representation for the word "apple" and "purple" is as shown in Figure 2.8, therefore, by adding these two vectors, the algorithm predicts the next word to be "plum". The reduced vector representation is what makes string vectors different from numeric vectors (Allahyari et al., 2017). The present research intends to adopt the principle of converting numeric vectors to string vectors in order to reduce dimensionality, and to eventually provide better predictability in classifying text.

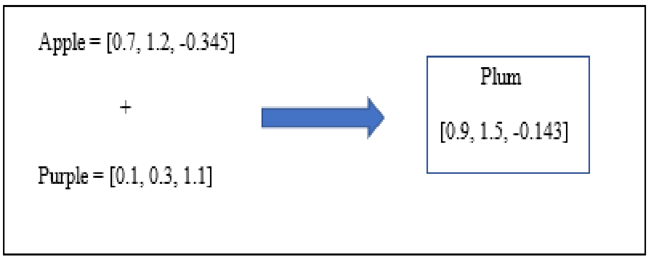


Figure 2.8 Vector representation example

2.11.3 Weighted Classification Algorithm using String Vector

This research applied string vectors as opposed to numeric vectors for classification, however, the semantic similarity between two string vectors is adopted from the cosine similarity between two numerical vectors as shown in Eq. 7. The advantage of using this method is a better classification results for the Multinomial Naïve Bayes classifier, as each string vector only has tens of dimensions compared to numerical vectors that have several hundred dimensions (Jo, 2016).

Past literature has shown major issues with numerical vectors include huge dimensionality and sparse distribution (Cosma & Acampora, 2016; Jo, 2016). Therefore, researches have adopted diverse preprocessing techniques to overcome these issues (Isguder-Sahin et al., 2014; Lochter, Zanetti, Reller, & Almeida, 2016; Natarajan et al., 2013; Uysal & Gunal, 2014).

Handling large feature sets for text classification is necessary, hence Harrag and Al-Qawasmah (2010) presented a number of dimensionality reduction techniques such as root-based stemming, light stemming and Singular Value Decomposition (SVD). Al-Anzi and AbuZeina (2017) also used SVD combined with Latent Semantic Indexing (LSI) to increase classification accuracy of Arabic text and to reduce dimensionality problems. In another study by Nassirtoussi, Aghabozorgi, Wah, and Ngo (2015), a multi-layer dimension reduction algorithm using semantics and sentiment was proposed to predict FOREX market exchange from news headlines. Nevertheless, the implementation of string vectors as an effort for dimensionality reduction seems to be limited. The most related work at implementing string vectors adopted the neural network models which are not only computationally expensive to implement but also require a very large amount of training data (W. Liu et al., 2017). Table 2.9 shows some of the related works improve classification and reduce dimensionality problems.

Table 2.9 Dimension Reduction Related Works

Reference	Features	Classifier	Dimensionality Reduction Technique	Model Adopted	Results
Al-Anzi and AbuZeina	TF-IDF	Support Vector Machine, Naïve Bayes, Logistic Regression	Singular Value Decomposition with Latent Semantic Indexing	Neural Network	SVM produced best result = 84.75%
Jo (2017a)	TF-IDF	k-Nearest Neighbor	String Vectors	Neural Network	Accuracy = 72%
Nassirtoussi et al. (2015)	TF-IDF	Support Vector Machine, k-Nearest Neighbor, Naïve Bayes	Semantics and sentiment	Own model	SVM produced best results = 82%
D. Zhang et al. (2015)	Word2vec	SVMperf	String vectors	Weka	Accuracy = 89.95%
Chatrath, Miao, Ramchand	Structured data	Logistic Regression	Structured data	Stepwise Multivariate Regression	Accuracy = 67%
Harrag and Al-Qawasmah (2010)	TF-IDF	Multilayer perceptron	Root based stemming, light stemming, singular value decomposition	Neural Network	Accuracy = 50%

2.12 Multi-Tier Classification Framework

As explained in Chapter 1, the term *multi* is defined as many or multiple indicating more than one while *tier* is defined as a row, rank or layer. For the purpose of definition within the context of this research, the term multi-tier will be used interchangeably with other similar terms such as multi-level and hierarchical. This section discusses studies in the past that have adopted a hierarchical classification framework.

Past research work looking into text classification has revolved around flat classification problems. Y. Liu, Bi, and Fan (2017) defined flat classification as the standard binary or multi-class classification problem where there are multiple features to fit the classification. However, literature has substantiated that flat classification would not be best suited to solve real-world problems as it is crucial for machines to think like humans in making decisions and predicting outcomes, hence this problem can only be addressed in a hierarchical classification framework (Silla & Freitas, 2011).

The availability of vast amount of online textual data has made it increasingly important to organize such documents hierarchically for better management. Research work in automatically classifying documents into prelabelled classes has shown there is a need to better organize such data before classifying them (Du et al., 2018). According to Baqapuri et al. (2016), classification performance is inversely proportional to the scalability of data available and the number of categories the data need to be classified into. In other words, classification time performance will suffer as the dataset gets larger. This however, can be curbed with the introduction of a hierarchical classification that arranges all categories into a tree-like structure and trains the classifier at each node of the given hierarchy (Du et al., 2018). The classification process takes place from the root of the tree, all the way down to its leaf node signifying the concluding category of the document. Du et al. (2018) further states how similar to flat classification, hierarchies are also illustrated as binary trees, however, each document is classified from the root of the tree but the direction of the tree is regulated by the respective node during a hierarchical classification process. The final classification category is determined by the leaf node of the tree. Figure 2.9 is a simple depiction of a hierarchical classification for diabetes.

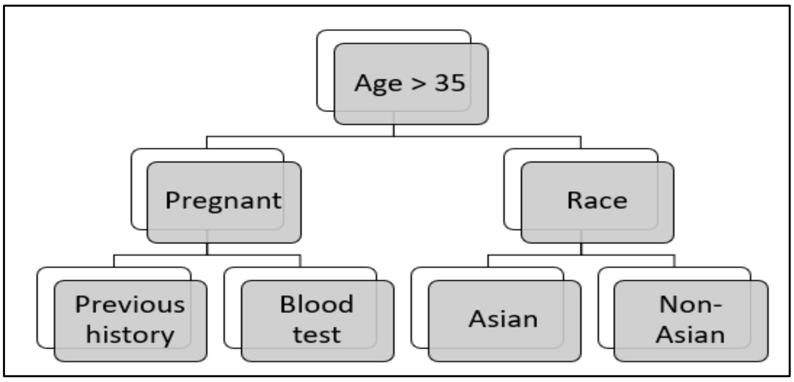


Figure 2.9 Sample hierarchical classification tree

In an exploratory study by McRoy et al. (2018), it was discovered that despite the abundance of information exchanged on online health forums, there still seems to be unmet information needs as the information extracted is disorganized, hence looking for what one needs becomes an arduous task. Other survey studies have also reported a proper health information organization to be beneficial not only to those who are actively posting within such communities but also to "lurkers" (i.e., those who read but do not post), as they keep returning over time to seek information when a need arises (X. Yang, Li, & Huang, 2017).

The literature work revolving adopting hierarchical classifications is very diverse from topic classification to image classification and text mining (Baqapuri et al., 2016; X. Fu, Li, Yang, Cui, & Yang, 2016; Mujtaba et al., 2017; Silla & Freitas, 2011). However, for the purpose of this research, the literature discussed within this section will focus on classifying text within the hierarchical structure.

Moh et al. (2015) o classified sentiments for a movie review dataset using four separate classifiers (Naïve Bayes, Support Vector Machine, Random Forest and Stochastic Gradient Descent) to experiment on the performance of the multi-tiered framework.. Their classifier detected positive, negative and neutral reviews in the first level, followed by classifying reviews according to polarity (more positive, less positive, more negative, less negative). The authors found that the Random Forest classifier outperformed the others, and a multi-tier framework was able to improve the prediction accuracy by more than 10% compared to a single-tiered framework.

In another study by Jinyan Li et al. (2016), a hierarchical filtering mechanism was applied using various classification algorithms (Naïve Bayes, C45, Decision Tree, Maximum Entropy, Winnow and Balanced Winnow) where the filtering mechanism progressively shrunk the dataset of online news articles with respect to contextual polarity and frequent terms of a document. In other words, the filtering system was done hierarchically where the first layer dealt with removing polarity words, followed by removal of high frequency words and unique high-frequency words. The authors found that Maximum Entropy outperformed all other classifiers used as it was able to incorporate different sources of information within its framework that helped overcome missing data.

With respect to the research work done in the past using a multi-tier classification framework, the attempt has only been made to a single element (either only sentiment classification, filtering mechanism, or emotion classification), with the number of tiers limited to two (Baqapuri et al., 2016; Ghazi, Inkpen, & Szpakowicz, 2010; S. Kim, Zhang, Chen, Oh, & Liu, 2013; Jinyan Li et al., 2016; Moh et al., 2015; Xu et al., 2015). The proposed framework for this research combines three distinct elements to be amalgamated into a single framework where posts will be classified according to type, sentiment, emotion and purpose. Table 2.10 shows some of the related research work adopting hierarchical frameworks.

Table 2.10 Studies Applying Hierarchical Classification Framework

Ref	Classification Type	Technique	Data Set	Results	Number of Tiers
Ive et al. (2018)	Text	Stochastic Gradient Descent with Hierarchical Recurrent Neural Network model	Mental health forums (Reddit)	Accuracy: 76%	2
Kowsari et al. (2017)	Document	Deep Learning, Recurrent Neural Network, Convolutional Neural Network, Hierarchical Deep Learning	WOS dataset	Accuracy: 90.93%	2
Mujtaba et al. (2017)	Document	Support Vector Machine, J48, Random Forest	Autopsy reports	Accuracy: SVM = 95.41% RF = 93.57% J48 = 88.99%	2
Jinyan Li et al. (2016)	Sentiment	Naïve Bayes, Decision Tree, C45, Maximum Entropy, Winnow, Winnow Balanced	Online news	Accuracy: SVM = 95.41% RF = 93.57% J48 = 88.99%	2
Moh et al. (2015)	Sentiment	Naïve Bayes, Support Vector Machine, Random Forest, Stochastic Gradient Descent	IMDB Movie Reviews	Accuracy: RF = 83.71% SGD = 82.19% SVM = 81.27% NB = 80.53%	2
Xu et al. (2015)	Emotion	SVR package in LibSVM	Sina Weibo	The lower layers of the hierarchy achieved higher accuracy	2
S. Kim et al. (2013)	Sentiment	Bayesian model with hierarchical scheme	Online product reviews	Accuracy: 85.7%	2

2.13 Summary

This chapter served as a purpose of literature studies conducted for the foundation of this research. Past studies that have looked to classify sentiment, type, emotion and purpose were looked into upon which gaps within the literature itself were identified.

In identifying techniques adopted for sentiment classification, it came to view that past literature looked more towards supervised machine learning algorithms for classification. And although the results were encouraging, there was a niche upon which had not been incorporated till date when it comes to classifying sentiment for Facebook posts which is leveraging on Facebook behaviors (like, share, comment and reaction). Literature has clearly shown how each behavior differs and how these behaviors impact on determining tickets pre-sales or disseminating information during a crisis. Yet, no study has been conducted to convert these behaviors as a contribution towards sentiment classification. This research looks to bridge this gap and proposed a formula converting the numbers into sentiment intensity which will be elaborated in the following chapter.

For type classification, it came to light that past studies within the diabetes domain has been very much focused on Type 1 diabetes or exploratory studies which include understanding the form of communication that takes place within social media platforms that caters to diabetes. Yet the studies for classification that has been carried out is filled with medical jargon extracted from hospital transcripts or doctor's notes. Additionally, a lexicon dictionary that caters to other form of diabetes extracted from social media (Type 2 and gestational diabetes) has not been made as yet, hence this is the next gap that this research can contribute to.

In emotion classification, past literature has been more focused on emotion detection rather than actual classification. When it comes to classifying emotion, the process seems to cater more towards positive emotions rather than negative emotions. This gap needs to be explored and methods to improve classification in this research are discussed in the following chapter.

Last but not least, to improve purpose classification, methods to convert numeric vectors to string vectors, using Multinomial Naïve Bayes was studied. Most studies in the past that have looked to adopt this model stayed within the limits of using artificial neural networks and studies that have adopted this model within a textual corpus produced very low accuracy due to dimensionality issues encountered. Past studies found numeric vectors tend to return good accuracy readings when it comes to classification, however the use of Word2Vec of converting numeric to string allowed the spaces within classification algorithm to overcome dimensionality issues and thus producing results with better accuracy readings. However, this concept has not been applied within textual data using semi-supervised machine learning approach hence this research looks to use the concept of numeric to string conversion for purpose classification.

Another contribution of this research is to classify posts according to type, sentiment, emotion and purpose in a multi-tier framework. Literature supports the idea of multi-tier frameworks being able to produce better classification, however, studies that have adopted such layers only did so for a single element of sentiment. This research looks to combine several elements of type and purpose that would eventually produce better sentiment and emotion classification.

CHAPTER 3: METHODOLOGY

Research methodology refers to the science of how a research has been carried out. The objective is to help readers understand the process involved in the research in order to reach the results. This chapter will look into the methodology adopted for this study. It will provide an in-depth elaboration on the steps taken in an effort to complete this research including a brief overview of the research architecture, data collection, data cleaning, experiments carried out etc.

The rest of this chapter is organized as follows: an overview of the operational framework and research framework is provided to familiarize readers with the flow of the research. This is followed by an in-depth look into each phase of the research which includes data preparation (from data cleaning to pre-processing). The proposed STEP framework is discussed (type, purpose, sentiment and emotion) next. Each tier is dissected and discussed in detail before moving on to the final phase of the research framework (i.e. evaluation set up) and finally the chapter is concluded.

3.1 Operational Framework

It is necessary to have a proper plan when conducting research. This helps in keeping the researcher focused and ensuring the objectives of the research are met. Figure 3.1 depicts the overall operational framework that has been adopted throughout this study.

The framework starts with the research initiation which includes literature review to identify gaps in previous studies, coming up with a problem statement as well as defining research objectives and scope. The next step is preparing the data which involves data collection, data clean up and pre-processing. This stage is necessary to ensure the data are ready for development phase. At the algorithm development phase, the algorithm goes through multiple cycles of testing and development until the results are satisfactory before evaluating the data using evaluation metrics as well as comparing it to other pre-existing frameworks.

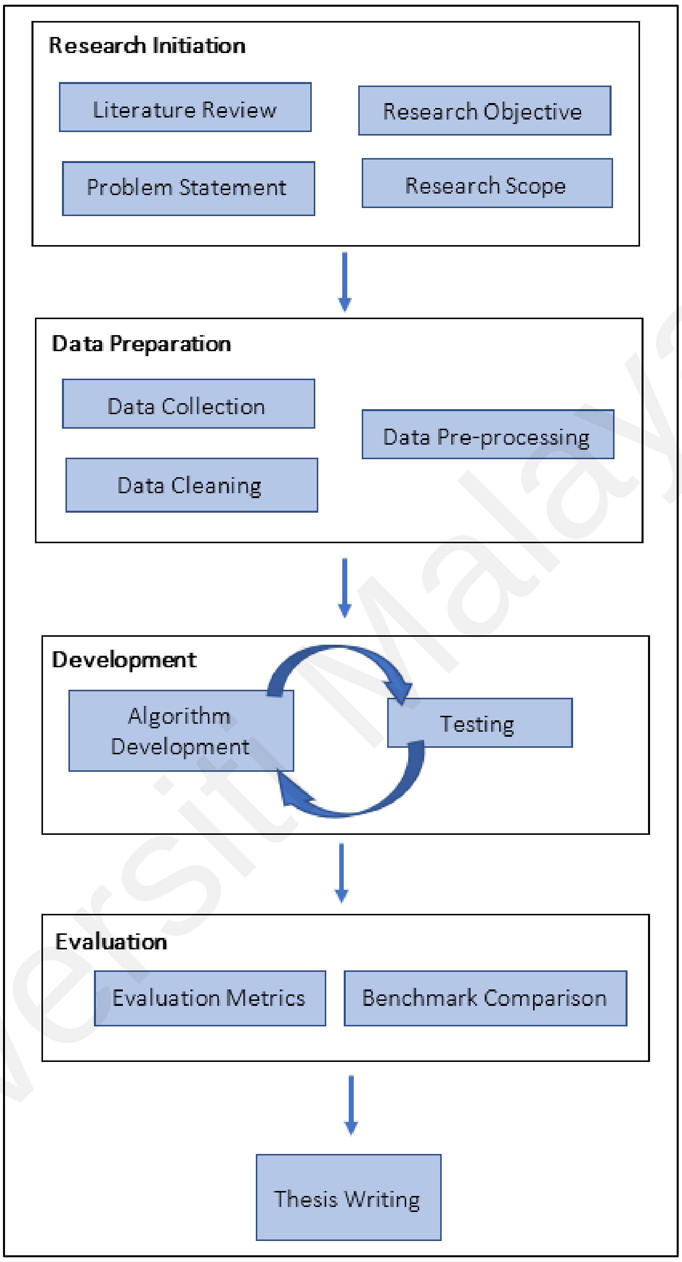


Figure 3.1 Operational Framework

3.2 Research Framework Overview

A general overview of the research framework is discussed within this section. This includes a brief introduction to the different phases involved in the framework. A detailed discussion on the proposed classification framework will be discussed in the upcoming sections.

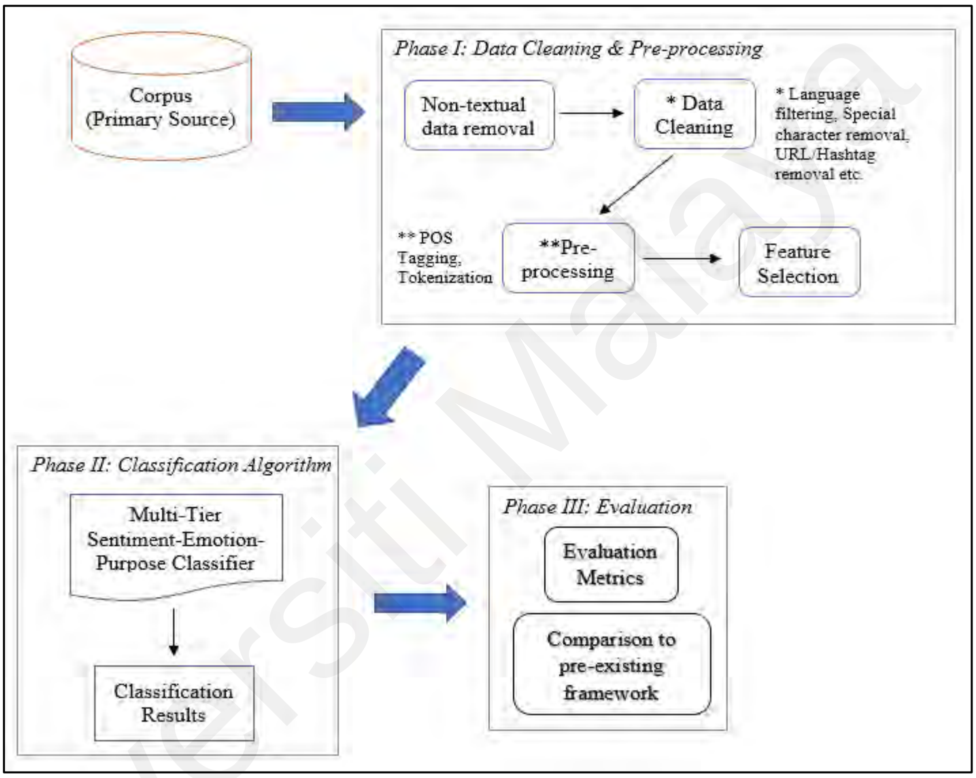


Figure 3.2 Research Overview

There were three phases involved in this research (Figure 3.2) namely, the data cleaning and pre-processing phase, building the proposed classification framework phase and finally evaluating the framework phase. Phase one looks into preparing the data for classification process; this includes removing non-textual elements, data cleaning and pre-processing. This is a crucial phase prior to the development of the algorithm (Haddi, Liu, & Shi, 2013; Singh & Kumari, 2016) and the pre-processing phase had to be visited many times to determine the final order of which pre-processing steps to adopt and which to omit. As literature has stated, different forms of data require different forms of pre-processing (Haddi et al., 2013). The data cleaning process includes language filtering, special character removal etc., while pre-processing looks into POS Tagging, stemming and tokenization etc. The following sections will describe each phase individually.

3.3 Phase I: Data Cleaning and Pre-processing

This section introduces the data source (groups from which data were extracted), the timeline for data extraction, data cleaning as well as the pre-processing mechanisms.

3.3.1 Data Source

Facebook has reported a total of 2.41 billion active users as of second quarter of 2019[11] and has been accepted as the most used social media platform compared to Instagram and Twitter (Groot, Westermann-Behaylo, Rehbein, & Perrault Crawford, 2019; Roopchund, Ramesh, & Jaunky, 2019). Therefore, the main data source for this research was Facebook. This was also due to the interest of the researcher to look into sentiment and emotion analyses of Facebook data considering most of the previous studies have focused on Twitter (Alahmari & Buckley, 2015; Alsaedi, Burnap, & Rana, 2017; Burnap et al., 2015) and online forums (Bu et al., 2016; McRoy et al., 2018; Sokolova & Bobicev, 2013) for sentiment or emotion analyses. Literature has also stated users looking for online help and support in combating diseases often turn to social media sites such as Facebook (Abedin et al., 2017; Gage-Bouchard et al., 2017; Harpel, 2018; Parackal, Parackal, Eusebius, & Mather, 2017).

As explained in Chapter 2, the subject matter for this research revolves around diabetes, hence, the keyword "diabetes" was used to search for groups and pages related

[11] https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

to this disease. Also, since one of the objectives of this research was to classify types of diabetes, the search criteria was specific to looking for top pages specifically catering to the different types of diabetes. Figures 3.3 shows a sample of one of the pages from which data were extracted:

Figure 3.3 Type 1 Diabetes Support & Information

(https://www.facebook.com/Type1DiabetesFacts/)

It was crucial to ensure the pages are active, therefore, all three pages (Table 3.1) were monitored and observed beginning July 2016. After following each page for 3 weeks, participants of the pages were found to be active in posting and responding to peers. The average posting was calculated based on the total number of posts collected within the 3 weeks divided by 21 days. The results were then rounded to the nearest number.

Table 3.1 Information on Pages/Groups

Group Name	No. of Participants	Average Posting During Observation Period	About Page
Type 2 Diabetes	81, 496	21	Started in 2014 for people affected by Type 2 Diabetes
Type 1 Diabetes Facts	77, 682	15	Founded in 2011 to educate the public on Type 1 Diabetes
Gestational Diabetes Information	5, 9007	12	Started in 2012 for women with gestational diabetes

3.3.2 Data Collection

Once the pages were finalized as the data source for this research, the data extraction process was carried out using Facebook Graph API[12]. This was done by first logging on to Graph API and creating an extraction app. An access token needs to be assigned to the extraction app in order to grant access to it to proceed for extraction of posts and comments. Besides posts and comments, the number of shares, likes and reactions was also extracted. Figure 3.4 shows sample data as extracted from Facebook.

Post Message	Love	Wow	Haha	Like	Sad	Angry	Thankful	Shares	Comments
For those who don't know.	4	2	0	259	20	2	0	79	30
A member of our community could use your help! Do children with insulin dependant diabetes need to follow a healthy eating plan ??	1	7	1	98	1	0	0	8	119
Supplements that could actually help with neuropathy? Read more about it	2	5	0	140	0	0	0	47	1
If you're planning to go on a diet, here are a few ideas that might help.	0	0	0	68	0	0	0	25	2
Here's how to properly dispose of syringes.	3	3	1	141	0	0	0	33	17
Yup.	12	9	108	720	54	4	0	302	58

Figure 3.4 Sample Posts Extracted

Data were collected for a duration of six months (July 2016 till January 2017). Total amount of data extracted from all three pages amounted to 78, 961 posts and comments.

[12]http://developers.facebook.com/docs/reference/api/

Prior to moving to the data cleaning and pre-processing phase, a total of 28,048 posts and comments were removed (i.e. 6,271 posts with emojis only, 9,919 spams and 11,858 posts with only the user names tagged), leaving 50, 913 posts to be carried forward to the next phase of this study. This is a crucial step to ensure all non-textual data are removed so that they do not interfere with the classification process (Akaichi et al., 2013; Toujani & Akaichi, 2017) .

3.3.3 Data Cleaning

Data extracted from any source of social media platform generally contains a lot of noisy data (i.e. negligible data). It is for this reason Singh and Kumari (2016), believe data cleaning is a vital step in preparing the extracted data for the classification process. The following elements were removed from the data collected:

a) Hashtags and URL links
b) Emoticons and emojis
c) Non-textual posts and comments (photo, video, GIF file etc.)
d) Posts and comments that were fewer than 3 words long
e) Posts and comments that were written in languages other than English
f) Posts and comments that had more than 5 misspelled words
g) Special characters (@, #, $, etc.)

Table 3.2 shows the justifications for removing the above-mentioned elements.

Table 3.2 Justifications for removal of elements

No	Element Removed	Justification
1	Hashtag and URL Links	These do not contribute towards classification process (N. Anand, Goyal, & Kumar, 2018; Krouska, Troussas, & Virvou, 2016)
2	Emoticons and emojis	Due to nature of text extracted, the number of emoticons and emojis used were too insignificant
3	Non-textual posts and comments	The scope of this research was to classify text for sentiment, type, emotion and purpose. Therefore, non-textual posts and comments were removed.
4	Posts and comments that were 3 words or less	During initial experiment phase, it was found the accuracy of sentiment, type, emotion and purpose classification drastically decreased for sentences that were too short. The cut-off was set to 3 words in this research
5	Non-English text	The number of English posts within the corpus was sufficient enough to ignore non-English text. This is also due to cost reduction in hiring a translator to translate every language spoken within the groups
6	More than 5 misspelled words	During initial experiment phase, it was discovered spell correction for all the texts consumed too much of pre-processing time, hence a decision was made to discard posts with more than 5 misspelled words
7	Special characters	These do not contribute towards classification process (N. Anand et al., 2018; Krouska et al., 2016)

Misspelled words are words that have been incorrectly spelt due to human error or typos such as *sometimes* spelt as *sumtimes* or *orally* spelt as *orraly* etc. For spell check purpose, the Wordnik API[13] dictionary (Python friendly API) was used. If a word was misspelled, it would automatically correct it but if a sentence contained more than five misspelled words, then the sentence will be discarded. The cleaning process resulted in a total count of 26, 531 raw data to work with (Table 3.3).

[13] https://github.com/wordnik/wordnik-python

Table 3.3 Distribution of data before and after cleaning

Group Name	Before Cleaning	After Cleaning
Type 2 Diabetes	20, 750	9, 891
Type 1 Diabetes Facts	19, 447	8, 626
Gestational Diabetes Information	10, 716	8, 014
Total posts	**50, 913**	**26, 531**

Additionally, posts and comments with more than 10,000 likes/shares/reactions were also discarded. This was decided upon after calculating the normal distribution bell curve to ensure final data are not skewed when proposing an intensity weight for Facebook behavior (likes, comments, shares, and reaction). Eventually, the number of posts came to 21, 082 posts. As mentioned in Table 3.3. the corpus comprised of posts and comments from three different types of diabetes, hence, to ensure an even distribution per diabetes type for training the algorithm (labelled data), 6, 000 posts (2, 000 posts per type) were randomly selected for human annotation. Pre-processing was then administered on these 6, 000 posts as described in the next sub-section.

3.3.4 Data Pre-processing

Data pre-processing refers to the process of preparing the extracted data for classification (Haddi et al., 2013; Singh & Kumari, 2016). In most cases, data extracted from social media are plagued by noisy and ambiguous parts such as scripts, HTML tags, special characters etc. Additionally, not all words in a sentence contribute towards classification. According to Haddi et al. (2013), the non-removal of words that do not impact the orientation of a sentence would only contribute towards a high dimensionality problem, and thus making classification more difficult.

The pre-processing steps include white space removal, abbreviation expansion, stemming, stop word and punctuation removal, tokenization and part of speech tagging (POS tagging).

3.3.4.1 Abbreviation expansion

An abbreviation dictionary consisting of common diabetes related terms was created based on simple observation of the data collected. The abbreviation list was created using an abbreviation dictionary obtained from the National Center of Biotechnology Technology[14]. The dictionary was created in Microsoft Excel. The full term of the abbreviation was replaced with a full version when encountered within the extracted data, however, if the abbreviation was not found in the dictionary, then the word was treated as a misspelled word and checked against Wordnik dictionary. Figure 3.5 is a sample of the dictionary created.

BDR	Background diabetic retinopathy
BFST	Behavioural family systems therapy
BMI	Body mass index
BMI-SDS	Body mass index – standard deviation score
BSPED	British Society for Paediatric Endocrinology and Diabetes
CASCADE	Child and Adult Structured Competencies Approach to Diabetes Education
CBT	Cognitive behavioural therapy
CGMS	Continuous glucose monitoring system
CHF	Chronic heart failure
CI	Confidence interval
CO2	Carbon dioxide
CSII	Continuous subcutaneous insulin infusion
CST	Coping Skills Training
CVD	Cardiovascular disease
DCCT	Diabetes Control and Complications Trial
DKA	Diabetic ketoacidosis

Figure 3.5 Sample Abbreviation Dictionary

[14] https://www.ncbi.nlm.nih.gov/books/NBK343415/

3.3.4.2 Tokenization

Tokenization is a process of splitting a given text into smaller fragments known as tokens (S. Sun et al., 2017). The Natural Language Toolkit 3.4 (NLTK) which is a suite for libraries for natural language processing for English texts was used for tokenization. Figure 3.6 shows a sample post and Figure 3.7 shows the corresponding tokenization.

Figure 3.6 Sample post

```
['When', 'people', 'with', 'diabetes', 'experience', 'a',
'dangerous', 'drop', 'in', 'blood', 'sugar', ',', 'glucose',
'tablets', 'might', 'be', 'a', 'better', 'option', 'than', 'a',
'sugary', 'food', 'or', 'drink']
```

Figure 3.7 Tokenized Sample

3.3.4.3 Stop Word and Punctuation Removal

Stop word removal is used to remove words that do not carry any impact on the orientation of a sentence and when removed, does not change the meaning of the sentence either (Jagtap & Adamuthe, 2018). Examples of stop words are ‘and’, ‘are’, ‘this’ etc. Punctuations were also removed along with the stop words. Figure 3.8 shows a sample data after stop words and punctuations were removed.

```
['people', 'diabetes', 'experience', 'dangerous', 'drop', 'blood',
'sugar', 'glucose', 'tablets', 'better', 'option', 'sugary', 'food',
'drink']
```

Figure 3.8 Sample data after stop words and punctuations were removed

3.3.4.4 Stemming

Stemming is used to remove suffixes or prefixes in order to get to the root of the word (S. Sun et al., 2017). For example, the words: *tokens*, *tokenizer*, *tokenization* and *tokenizing* can be reduced to the root word *tokenize*. From Figure 3.8, the only word that would require stemming would be *tablets* which would be stemmed to the word *tablet*.

3.3.4.5 Part of Speech Tagging (POS Tagging)

Part of speech tagging refers to the practice of corresponding words in a text to a particular part of speech (S. Sun et al., 2017). In other words, it is simply identifying each word to its grammatical category of nouns, verbs, adjectives and adverbs etc. Figure 3.9 shows the corresponding POS tags for the sample data where NN refers to a noun in a singular form, JJ means adjective and JJR refers to a comparative adjective.

```
['people' [NN], 'diabetes' [NN], 'experience' [NN], 'dangerous'
[JJ], 'drop' [NN], 'blood' [NN], 'sugar' [NN], 'glucose' [NN],
'tablet' [NN], 'better' [JJR], 'option' [NN], 'sugary' [JJ], 'food'
[NN], 'drink' [NN]]
```

Figure 3.9 Corresponding POS tags

3.4 Phase II: Multi-Tier Sentiment-Type-Emotion-Purpose (STEP) Classifier

This section introduces the proposed multi-tier sentiment, type, emotion and purpose (STEP) classifier, including the preliminary experiments conducted followed by an in-depth look at the methodologies adopted within each tier of the STEP classifier.

Out of the 6, 000 posts, the percentage decided upon to split the data was set to 30% for testing and 70% to train the algorithm. Literature has claimed semi-supervised text classification requires a larger amount of training data compared to supervised (da Silva et al., 2016; Fernández-Gavilanes et al., 2016; García-Pablos et al., 2018; Lo, Cambria, Chiong, & Cornforth, 2016), therefore the decision to split the data accordingly was justified.

3.4.1 Human Annotation

A total of 6, 000 posts were sent to thirteen experts for annotation, deemed necessary to prepare the labelled dataset. These experts comprised of three linguists, seven medical experts and three researchers. The medical experts were three doctors and three nurses from the pediatric diabetes ward of Sungai Buloh hospital and one from the Faculty of Medicine, University of Malaya. The linguists and researchers were from the Faculty of Languages and Linguistics and Faculty of Computer Science and Information Technology, University of Malaya respectively.

Each expert was given approximately thirty days to complete the annotation. However, only responses from eleven experts were received. At this junction, the inter-rater reliability (IRR) was determined whereby the Krippendorff alpha was calculated using IBM SPSS statistical analysis software. Krippendorff alpha coefficient is a statistical calculation of the agreement used for analysis where textual units are converted to analyzable terms often called the inter-coder agreement or inter-rater reliability

(Krippendorff, 2004). Burnap et al. (2015) states Krippendorff alpha is statistically more dependable as it evaluates the degree of agreement between each annotator by calculating the disagreement range opposed to the agreement range. The basic formula for Krippendorff alpha is simple yet it involves complex resampling computational methods (Krippendorff, 2011) where the results range between 0 (perfect disagreement) to 1 (perfect agreement). Krippendorff (2004) states the inter-rater reliability should be $\alpha \geq 0.800$, however, the lowest acceptable limit is $\alpha \geq .667$. For this research, the percentage achieved was $\alpha = 0.89$ (89%), indicating the percentage of disagreement between the annotators were insignificant, and thus the result of the annotation is statistically accepted.

3.4.2 Preliminary Experiment

Prior to proposing a classifier combining all four elements within a single framework (sentiment, type, emotion and purpose), multiple experiments were conducted. Some of the results proved fruitful to be adopted while others served as a learning lesson. This sub-section describes each of the preliminary experiment.

3.4.2.1 Topic Modelling Experiment

As explained above, there are four tiers to the proposed STEP framework. Classifying diabetes type is straight forward as there is a distinction between each type (type 1, type 2 and gestational diabetes). This will be discussed in the sections that follow. Same is applicable to sentiment (positive, negative, neutral) and emotion (anger, joy, trust, disgust, fear etc.) classification as the groups are already predetermined. However, when it came to classifying for purpose, the idea was to generate the groups from within the corpus itself where a topic modelling algorithm could be used to identify the groups within the corpus without having the need of human intervention to scrape through the data at hand to determine groups manually.

With the sheer amount of text available for analysis, the initial idea was to find common features between them in order to begin purpose classification. It was crucial to determine the major topics being discussed within the diabetes groups in order to identify the purposes. Literature has identified several reasons or motives such as providing support, seeking treatment information and sharing personal experiences (Yan et al., 2016). However, these were very broad purposes which could eventually be narrowed down to specific purposes. Furthermore, the needs of groups battling one disease may differ from another (i.e. need for bereavement support in case of cancer patients compared to patients battling diabetes). In order to understand the data at hand, one of the methods adopted was to run a topic modelling experiment.

Topic modelling or topic classification as it is used interchangeably (S. Sun et al., 2017), is a form of unsupervised learning akin to clustering, where the set of topics being discussed are unknown prior (Hashimoto, Kontonatsios, Miwa, & Ananiadou, 2016; K. W. Lim, Buntine, Chen, & Du, 2016). According to E. H.-J. Kim, Jeong, Kim, Kang, and Song (2016), a topic is merely a repeated pattern of co-existing terms in a corpus. Therefore, a good topic model for *healthcare* should result in terms like *health*, *doctor*, *patient*, *hospital* etc. while for *agriculture* it should result in terms such as *crops*, *farm*, *wheat* etc.

The present study adopted the most common topic modelling approach known as the latent Dirichlet allocation (LDA) (Al-Anzi & AbuZeina, 2017; Calheiros, Moro, & Rita, 2017; S. Lim et al., 2017; Yoon, Kim, Kim, & Song, 2016). The training process used the prebuilt *topicmodels* package with full dependencies and returned after 2000 iterations of Gibbs sampling, with k = 50 topics, and Dirichlet hyper-parameters $\beta = 0.1$ and $\alpha = 50 / K$

LDA assumes the documents (corpus) presented as an input as a mixture of topics. Those topics would further generate words based on a probability distribution (Calheiros et al., 2017). Hence, given a set of documents, LDA works by backtracking those documents to evaluate topics that created those documents to begin with. Figure 3.10 shows the model adopted for this preliminary experiment of topic modelling.

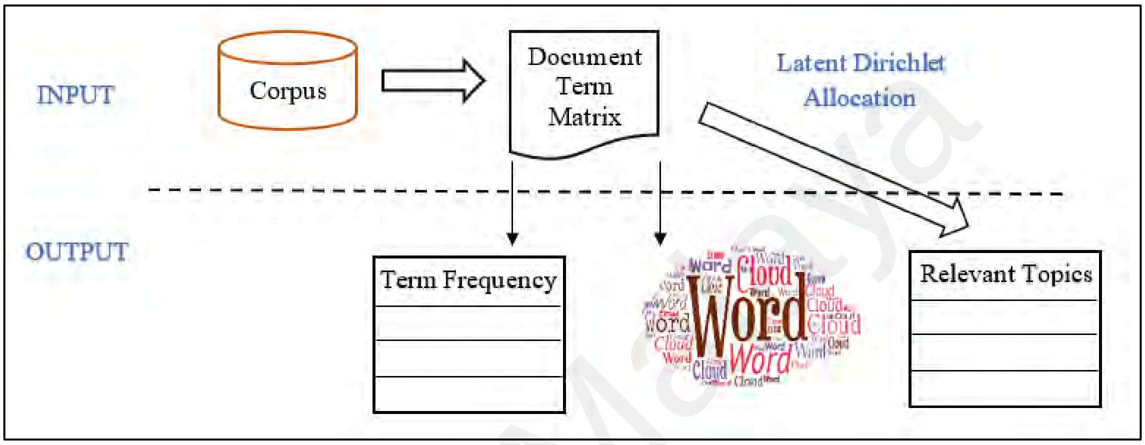


Figure 3.10 LDA Approach for Topic Modelling (Calheiros et al., 2017)

There are three key steps involved in performing LDA; cleaning and pre-processing, preparing document term matrix and running the LDA model. The three outputs expected are the term frequency, word cloud and a list of relevant topics. The coding for this stage was done using Python 3.7 and the corpus was converted to a document term matrix using the built-in library “genism”. Figure 3.11 shows the pseudocode used to run the LDA model.

ALGORITHM: *Topic_Modelling*

Input: Corpus of Facebook posts
Output: Topics from posts

Procedure

1. Dataset is loaded
2. Dataset is pre-processed
3. Bigrams and trigrams created from dataset
4. Build LDA Topic Model
5. View model

End

Figure 3.11 Pseudocode to run LDA model

The document term matrix produced a term frequency table listing words with the highest frequency, which was then visually presented in the form of a world cloud. Hence, words with the highest frequency were shown with the largest font and gradually decreasing in size as the frequency decreases.

In order to execute an LDA model, the number of topics is a required parameter. Calheiros et al. (2017) found the number of topics can be fine-tuned in order to produce ideal results. Therefore, the initial number of topics in this experiment was set to twenty-five (i.e. 50% of the total number of term frequency produced by the document term matrix). This resulted in a few coinciding topics which implied the optimal number of topics serving as a parameter for the model should be lesser than twenty-five. Therefore, the process was repeated, and the ideal number of topics was set to fifteen.

The topics produced by the LDA model needed human intervention in order for them to be grouped accordingly. Figure 3.12 shows a sample of the groups proposed with respect to the topics. Further details will be provided in section 3.4.3.2.

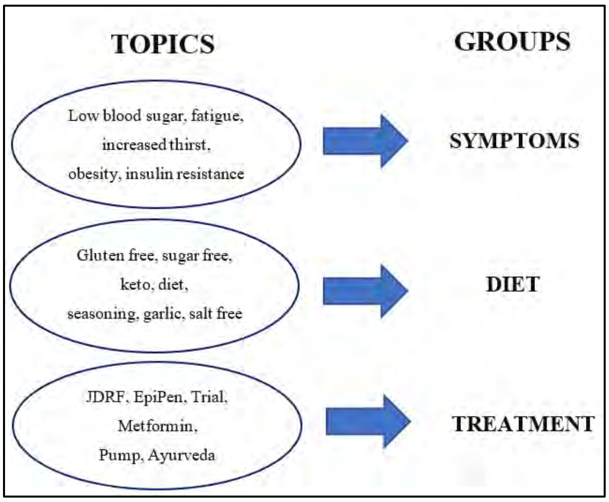


Figure 3.12 Inferring Topic from Keywords

3.4.3 Multi-Tier Sentiment-Type-Emotion-Purpose (STEP) Classification Framework

The proposed classification framework comprises of four tiers; sentiment, type, emotion and purpose (Figure 3.13). Each tier will be discussed separately in the sub-sections ahead.

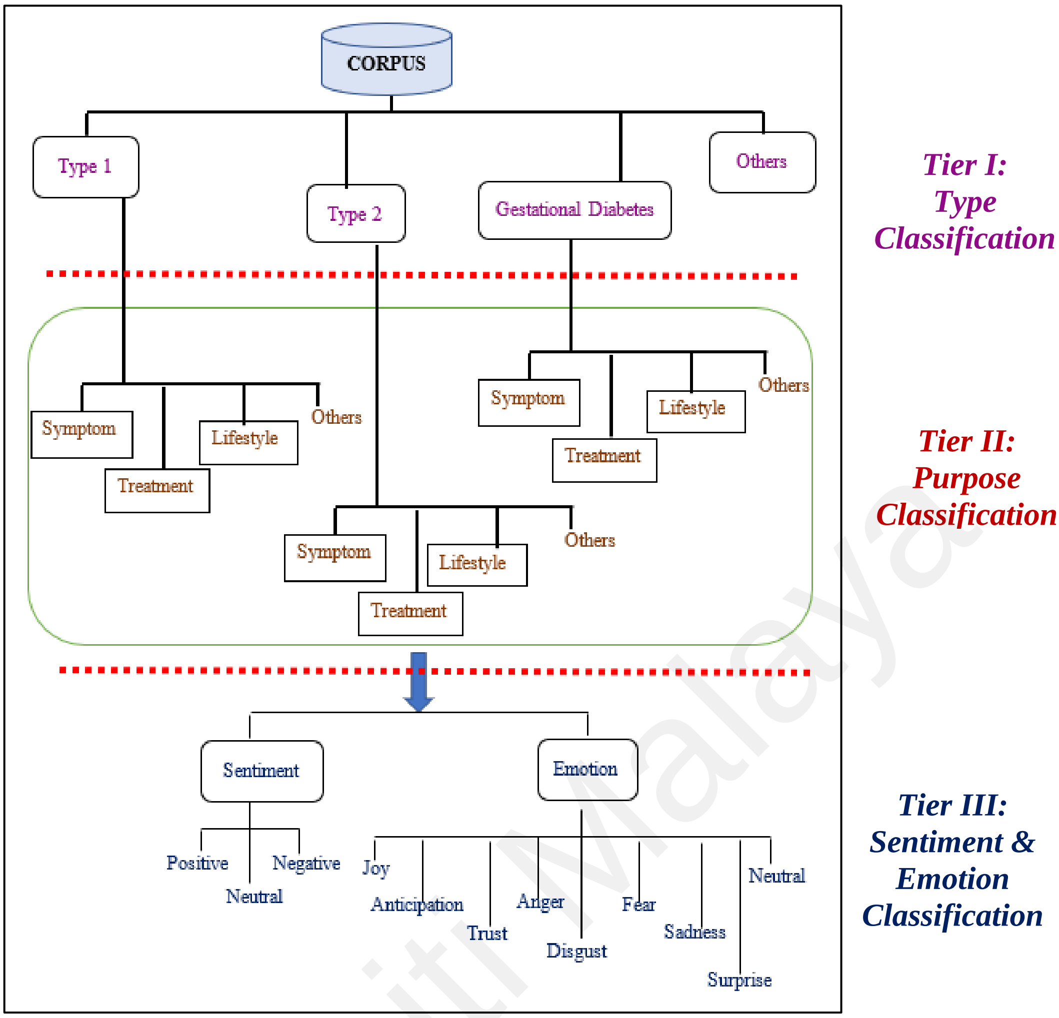


Figure 3.13 Multi-tier STEP Classification Framework

3.4.3.1 Tier I: Type Classification

Since there were no lexicon dictionaries for all three diabetes types readily available that could help classify the tokens into the correct categories, manual diabetes lexicon dictionaries were created. This is the main contribution made within this tier; i.e. the creation of lexicon dictionaries that cater for all three types of diabetes. Figure 3.14 shows a sample of lexicon for type 1 diabetes. Likewise, two other lexicon dictionaries catering for type 2 and gestational diabetes were also manually created.

type 1	hormone insulin
ketoacidosis	insulin dependent
insulinoma	insulin resistance
pump	high fat
dialysis	glucose monitoring
hypoglycemia	glycated haemoglobin
hyperglycemia	extreme thrist
glucagon	frequent urination
lipohyertrophy	vaginal infection
epipen	bacterial infection
nephropathy	
neuropathy	

Figure 3.14 Sample STEP Classifier Type 1 dictionary

A total of 1000 lexicons were given to the medical experts (Section 3.4.1) for verification, of which 780 were approved. The final lexicon list also comprised of a feature list of unigram and bigrams generated from the corpus itself to aid the diabetes type classification. The inclusion of unigram and bigrams was determined after conducting experimental runs using the diabetes dataset, the results of which will be presented in Chapter 4.

Naïve Bayes is a well-known classifier that can be easily implemented using scikit-learn from NLTK. Apart from that, classifying type was proven to be a binary classification problem. Literature has shown that Naïve Bayes yields the best results in binary classifications (Barnaghi et al., 2016; Nayak, Pai, & Pai, 2016). Therefore, for this tier, the Naïve Bayes classifier was adopted. It is to note that a post that could not be classified into a type was classified as *Other*. Figure 3.15 shows the pseudocode for this tier.

```
ALGORITHM: Type_Classification

Input: Training set, Testing set,
Dictionary of Type1 keywords (t1 f),
Dictionary of Type2 keywords (t2 f),
Dictionary of Gestational Diabetes (t3 f)
Output: Type classification of posts
Notation: FB: corpus, S: sentence, T F: feature file,
T: feature (unigram, bigrams), Training_set: train_x,
Testing_set: test_y

Procedure
while S in FB do
    T F = get_Tokens (S)
    for T in T F do
        if (T is present in t1 f) then
            Type of S is Type1
        end
        elseif (T is present in t2 f) then
            Type of S is Type2
        end
        elseif (T is present in t3 f) then
            Type of S is Gestational Diabetes
        end
        else
            Sentence is treated as type Other
        end
    end
end
Split data to train_x, test_y
Apply Logistic Regression Model
Evaluate model
End
```

Figure 3.15 Type Classification Pseudocode

3.4.3.2 Tier II: Purpose Classification

One of the contributions of this research as mentioned in Chapter 1 lies within this tier. This tier aims to classify Facebook posts that have already been classified into types in the previous tier into different purposes. According to S. M. Mohammad et al. (2015), purpose refers to a reason a statement is made. In the context of this research however, purpose is used as the topic upon which posts are classified. Simply put, purpose classification categorizes posts according to pre-determined classes.

The first step in this tier is to determine the number of classes based on the numbers of topic identified via the topic modelling experiment . The number of classes was initially identified to be ten (cause, symptom, exercise, modern treatment, traditional treatment, emotional support, financial aid, advice and promotion), however, test runs yielded inconsistent results. Furthermore, data for some classes such as financial aid were lacking from the extracted corpus (less than 50 posts), hence such classes were excluded.

Similarly, within the advice and emotional support categories, there were features that were identified relating to either discussing diet plans or treatment options. It also became apparent that classes such as modern and traditional treatment could actually be categorized into the main theme of Treatment. As such, with the help of linguistic and medical experts, the proposed classes from the topic modelling experiment were refined, resulting in four classes namely, Symptom, Lifestyle, Treatment and Other. The explanation of each feature class can be seen in Table 3.5 below.

Table 3.4 Features identified for purpose classification

Name of Class	Feature Identified
Symptom	Symptoms of each type of diabetes As an example, this includes extreme thirst and frequent urination for type 1, hormonal imbalance for type 2 and higher level of fat in the abdominal area for gestational diabetes Some of the symptom classification is also based on the reaction towards medication or a diet.
Lifestyle	Lifestyle covers both exercise options and diet options for patients battling diabetes. Posts classified here also include tried and tested recipes as well as diet programs that have worked well for other patients. Dietician and nutrition advice are also categorized within this class
Treatment	Both modern and traditional medicines are classified within this group. This includes discussions on the prices of medication and clinical trials to combat diabetes.
Other	All posts that do not have a distinct classification into either one of the classes stated above are grouped here.

It is to note that some labels overlap with one another. For example, the sample sentence below shows how the **treatment** drug (Metformin) is causing the patient to encounter nauseous like **symptoms**:

I've been on ***Metformin*** *for a month now but I can't stand the side effects of it. I constantly feel* ***nauseous****.*

With respect to the sample above, it was discovered the data was catering more towards multi-label classification problem, thus co-training Multinomial Naïve Bayes algorithm with weights was used in building the purpose classifier. C.-H. Lee (2018) stated in his study that multi-label classification documents provide a natural environment for co-training. Hence, similar to the methodology adopted by C.-H. Lee (2018), the features identified for purpose classification were also divided into two views: the label value set (LVS) which are more traditional features such as TF-IDF, Chi Square and Mutual Information Gain, to name a few and word feature set (WFS) which are class labels transformed to binary labels. In this research however, an attempt was made to add weights to LVS and convert WFS into string vectors in order to reduce dimensionality as discussed in chapter 2. The definitions of both LVS and WFS are as stated below (C.-H. Lee, 2018):

Definition of LVS:

Assuming c_i as the *i-th* class value,

$$\text{LVS} = \{\, c_i \mid i = 1, 2, \ldots, L\}$$

Where L = class label

Using the above definition, each label is treated as a binary feature. For instance, assuming there are five possible labels and the post is multi classified as c_2, c_3 and c_4. Hence, the multi-class value set of $\{c_2, c_3, c_4\}$ is represented as $\{0, 1, 1, 1, 0\}$.

Definition of WFS:

Assuming a_i is the word frequency bin for the *i-th* word,

$$\text{WFS} = \{\, a_i, c_j \mid \forall i, j = 1, 2, \ldots., L\} = \{\, a_i \mid \forall i\} \cup LVS$$

From the above definition, it can be seen that WFS is a culmination of word frequency bin (deduced from frequency of words within a document) and the LVS itself. Here, the frequency term is converted into word frequency bins. At this stage, weighting methods proposed within this research were applied to Multinomial Naïve Bayes.

In the co-training algorithm adopted from C.-H. Lee (2018), two base classifiers were used; Dependency Classifier (DC) which is associated with LVS and Feature Classifier (FC) which is associated with WFS. In DC, labels are regarded as independent features thus the classification process is carried out using only the label information where LVS is used both as input and target features.

The changes made to DC at this point of the research was to add weights on the mutual gain information algorithm used to identify features. The reason mutual gain was used instead of other feature selection techniques was its ability to identify terms that were identical within multiple labels. Moreover, it also works well with Multinomial Naïve Bayes (Ayoub Bagheri, Saraee, & de Jong, 2013; Bravo-Marquez et al., 2016; C.-H. Lee, 2018). Figure 3.16 shows the pseudocode for the weighted information gain function adopted at the DC level.

```
ALGORITHM: Weighted Mutual Information Gain
Input: Training set with observations X and corresponding
       labels Y
Output: feature set S of DC
Notation: DC: Dependent_Classifier

S ← Ø;
W ← 1 {Same weight for all samples}
while stopping criterion not true do
  F_max = arg max [wI(F_i,T,W)] {Find feature with maximum weighted MI}
  S ← S ∪ F_max {Add feature to subset}
  F ← F \ F_max {Remove feature from candidate set}
  Classifier ← Train DC Classifier (X, S, Y)
  Y′ ← ApplyClassifier (Classifier, X)
  W ← |Y - Y′| {Residual of each sample is new weight}
  CheckStoppingCriterion ()
end while
```

Figure 3.16 Weighted Information Gain Pseudocode

The fundamental idea behind the weighted feature is that in order to identify features to be included to improve the classification process, the focus should be on the mistakes made instead of the whole corpus. Therefore, correctly classified features were excluded and only those samples that have been misclassified were then assigned an equal weight. This residual weighting adjustment continues for each sample in the subset.

The results from the DC classifier were then fed into the feature classifier (FC) which is built upon the value weighted Multinomial Naïve Bayes, where for each class label c_l, it predicts the relevance of the *i-th* label using WFS. As explained earlier, WFS contains both class labels and regular feature words rendering FC as a multiplicative form of both components. The prediction of DC is automatically forwarded to FC in order to obtain the final classification. In this research, the binary classification of class labels represented as 1 and 0 were converted to string vectors to reduce the number of features. Figure 3.17 below shows the pseudocode for the multi-label co-training weighted algorithm using string vectors.

ALGORITHM: *Purpose_Classification*

Input: Training set, Testing set, Predicted label set
Output: Predicted label set according to purpose
Notation: *DC:* Dependent_Classifier, *FC:* Feature_Classifier
$\hat{C}$: *feature_label*

while $\hat{C}$ **do**

1. Learn DC and predict class value
2. Learn FC and predict class value

end

Figure 3.17 Purpose Classification Pseudocode

3.4.3.3 Tier III: Sentiment Classification

Once the posts were classified according to purpose, each of the purpose was then classified according to sentiment (positive, neutral and negative). In order to conduct sentiment analysis, each post had to be treated as a tuple. The tuple for this research was inspired from Shamim (2015), where each post contained two main elements: metadata (MD) and body (B). The metadata contained the number of likes, comments, shares and reactions while the body consisted of sentences. Each sentence tuple additionally comprised of three elements [S, I, C] where S represents the semantic polarity (positive, negative or neutral), I refers to the intensity of the given semantic and C showcases the content of the post. Figure 3.18 is a graphical representation of the tuple.

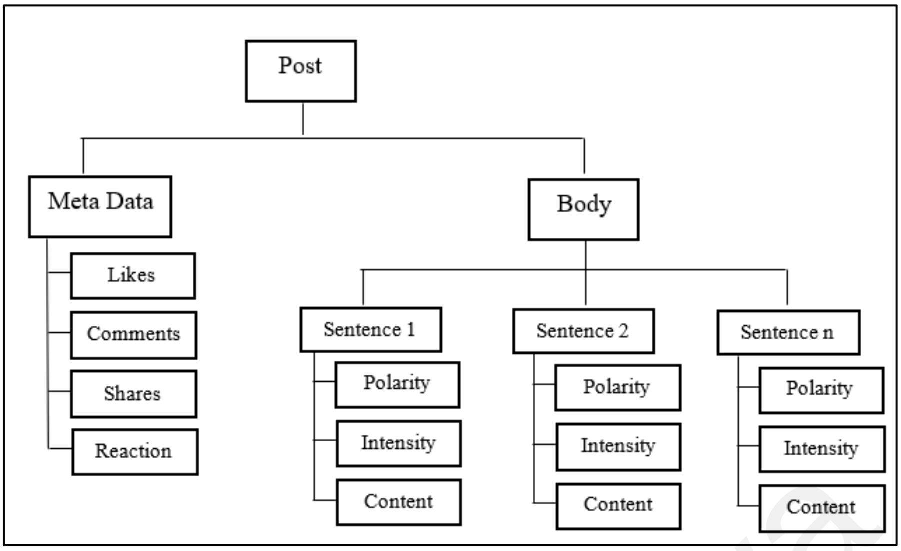


Figure 3.18 Graphical Representation of a sentiment tuple

SentiWordNet 3.0 (Baccianella et al., 2010) was used as the lexicon dictionary to determine the sentiment polarity.. The polarity ranges from +1 to +5 (where +1 indicates mildly positive and +5 as strongly positive) for positive sentiment and -1 to -5 (where -1 indicates mildly negative and -5 indicates strongly negative) for negative sentiment. The dictionary worked by calculating the terms of sentiment bearing words and assigning a sentiment score for the whole post.

However, literature has showed other elements that implicitly contribute to the sentiment of a posts, especially with the existence of Facebook behaviors such as likes, comments, shares and reactions (Calero, 2013; Quesenberry & Coolsen, 2018; Zell & Moeller, 2018). Facebook EdgeRank[15] algorithm places more weight on the number of *shares* compared to *likes* and *comments*, however, the number of *comments* outweighs the numbers of *likes*. Based on this acumen, the weight of these behaviors is an indirect expression of sentiment as users tend to either *share*, *like* or compelled to *comment* when they are in agreement with a post (C. Kim & Yang, 2017). In order to formulate a feasible

[15] http://edgerank.net/#How-does-EdgeRank-work

weight per behavior, multiple experiments were conducted using different weights. Although a higher number of such behaviors encountered per post indicate a stronger sentiment, however it does not indicate the absolute strength of the sentiment, i.e. a thousand *shares* or *comments* would imply more intense sentiment, but it does not mean it's a thousand times as much. Table 3.6 shows the final weights assigned to each behavior for this research.

Table 3.5 Behavior weights

Behavior	Weight	Range
Like	0.05	Per 1000
Comment	0.1	Per 1000
Share	0.2	Per 1000
Positive Reaction	+0.5	No range
Negative Reaction	-0.5	

From the table it can be seen for each thousand of *likes* for example, the weight assigned is 0.05. For the next thousand, this weight will increase to 0.1 and so on. The same calculation applies to *comment* and share, yet *reaction* has a different range. For positive reaction (love, wow, haha) the weight assigned is +0.5 while -0.5 is assigned for negative reaction (sad, angry). Furthermore, the weight for *reaction* is set at 0.5 despite the number of *reactions* encountered, unlike other behaviors as the literature has shown that the use of likes, comment and shares have more significance in comparison to reactions (C. Kim & Yang, 2017; Quesenberry & Coolsen, 2018; Ye Tian et al., 2017; Zell & Moeller, 2018). The weight of 0.5 has been set after conducting experiments within the dataset to determine the proper weightage. The formula is also dependent on the sentiment score obtained from SentiWordNet (SentiWordNet score is added to sentiment intensity to determine final sentiment score) as in if the score obtained from SentiWordNet is negative, then the formula will carry the negative symbol for its weightage and vice versa if the score is a positive. Eq. 8 below shows the proposed equation to convert the behaviors extracted into intensity for sentiment classification.

$$\sum_{i=1}^{n} \log\big((number\ of\ \alpha) + 1\big)\ x\ (\pm weight\ of\ \alpha) \qquad \text{Eq. (8)}$$

Where n = number of behaviors,

α = behavior (likes, shares, reaction, comments)

As an example, consider the sample post (Figure 3.19) and its breakdown of the behaviors (Table 3.6):

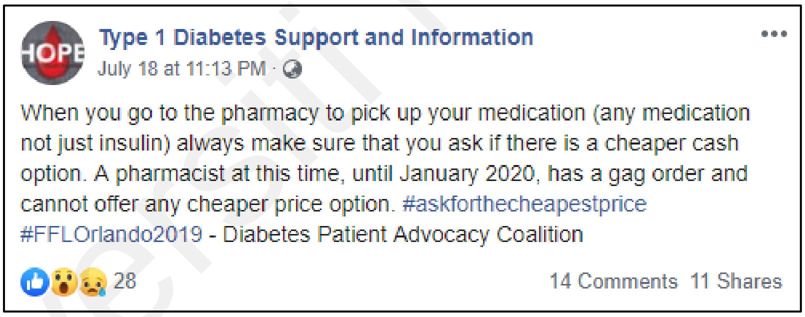


Figure 3.19 Sample negative post

Table 3.6 Behavioral Data Collected from Sample Post

Love	Wow	Haha	Like	Sad	Angry	Share	Comment
0	3	0	22	3	0	11	14

The SentiWordNet score for Figure 3.19 was -0.135, indicating a negative sentiment. Therefore, all the assigned weights for the behaviors would be negative (reaction is by default negative considering the use of sad).

Therefore, the sentiment intensity calculation would be as follow:

Like = (log(22) + 1) x (-0.1) = -0.234

Share = (log(11) + 1) x (-0.2) = -0.408

Comment = (log(14) + 1) x (-0.05) = -0.107

Reaction (Sad) = (log(3) + 1) x (-0.5) = -0.739

Therefore, final score = -0.135 - 0.234 - 0.408 – 0.739

= -1.516

Figure 3.20 below shows a similar example for a positive post. Table 3.7 shows the corresponding behavioral data.

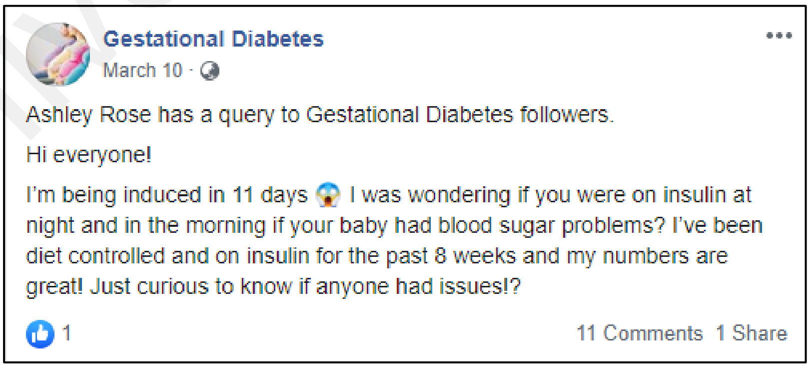


Figure 3.20 Sample positive post

Table 3.7 Data Collected from Sample Post

Love	Wow	Haha	Like	Sad	Angry	Share	Comment
0	0	0	1	0	0	1	11

The SentiWordNet score for Figure 3.20 was 0.098, indicating a positive sentiment. Therefore, all the weights remain to be positive including reaction (i.e. no negative reaction for the sample post as shown in Table 3.7).

Therefore, the sentiment intensity calculation would be as follow:

Like = (log(1) + 1) x 0.1 = 0.1

Share = (log(1) + 1) x 0.2 = 0.2

Comment = (log(11) + 1) x 0.05 = 0.102

Therefore, final score = 0.098 + 0.1 + 0.2 + 0.102

= 0.5

The calculation of sentiment polarity with added intensity using the formula proposed is depicted in the pseudocode provided in Figure 3.21

ALGORITHM: *Sentiment_Classification*

Input: Corpus from Purpose Tier
Like count (Like_C), Comment count (Comment_C),
Share count (Share_C), Reaction_Pos (Pos_C), Reaction_Neg (Neg_C)
Output: Sentiment classification of posts
Notation: *C*: corpus, *S*: sentence

```
while S in C do
    count = 0
    Sent_S = Sentiment_Score (S)
    if (Sent_S > 0.000) then
        if (Like_C > 0) then
          Calculate like score use Eq. 8
           count = Like_Score
        end if
        if (Comment_C > 0) then
          Calculate comment score use Eq. 8
           count = count + Comment_Score
        end if
        if (Share_C > 0) then
          Calculate share score use Eq. 8
           count = count + Share_Score
        end if
        if (Pos_C > 0) then
          Calculate like score use Eq. 8
           count = count + Reaction_Score
        end if
    end
    else
        if (Like_C > 0) then
          Calculate like score use Eq. 8 * (-1)
           count = Like_Score
        end if
        if (Comment_C > 0) then
          Calculate comment score use Eq. 8 * (-1)
           count = count + Comment_Score
        end if
        if (Share_C > 0) then
          Calculate share score use Eq. 8 * (-1)
           count = count + Share_Score
        end if
        if (Neg_C > 0) then
          Calculate like score use Eq. 8
           count = count + Reaction_Score
        end if
    end
    Sentiment_Score = Sent_S + count
end
```

Figure 3.21 Sentiment Classification Pseudocode

3.4.3.4 Tier III: Emotion Classification

The final tier of the proposed STEP framework is emotion detection. NRC Word-Emotion Association Lexicon (Emolex) (S. M. Mohammad & Turney, 2013) was used for emotion detection in this tier. Emolex is a lexicon collection of English words linked to emotional vectors with reference to Plutchik's eight basic emotions (Plutchik, 2003) namely; anger, fear, anticipation, trust, surprise, sadness, joy, and disgust and two sentiments (negative and positive).. As the upper classification tier has already separated sentiment to positive, negative and neutral; this final tier focuses on detecting emotions from the given sentiments.

With respect to the literature discussed in Chapter 2, one of the main area of concerns in emotion detection is the ability to correctly identify the emotion from a text. Canales, Strapparava, Boldrini, and Martnez-Barco (2016) discovered using readily available emotion lexicons such as Emolex, EmoSenticNet or DeepcheMood as is may not be appropriate as each dataset is very subjective, hence words used in a different context would carry different emotions. For example:

The numbers are up!

The above could either indicate *joy*, *sadness* or *fear* depending on the context of the sentence. In case of a finance corpus, it would translate to *joy* but if it was related to medical diabetic patients, it indicates *fear*. Therefore, the effort of this research is to adopt bootstrapping method to automatically annotate emotional corpora within the context of the diabetes corpus in order to accurately detect emotions. This method was inspired from Canales et al. (2016). Figure 3.22 shows the pseudocode adopted for emotion detection using Emolex. The vector assignment shown as [0,0,0,0,0,0,0,0] (Figure 3.22) shows the emotion assignment of Emolex for each emotion, i.e. anger, fear, anticipation, trust, surprise, sadness, joy, and disgust, respectively.

```
ALGORITHM: Emotion_Vector_Assignment

Input: Corpus from Sentiment Tier
Output: Emotion detection of posts
Notation: C: corpus, S: sentence, E: Emolex
T F: feature file, T: feature (unigram, bigrams)

while S in C do
    emotion_vector = 0
    for T in T F do
        if (T exist in E) then
          emotion_vector = Emolex_vector + emotion_vector
        end if
        count emotion_vector
        assign emotion in E
    end
    sort emotion_vector
end
```

Figure 3.22 Emotion Vector Assignment Pseudocode

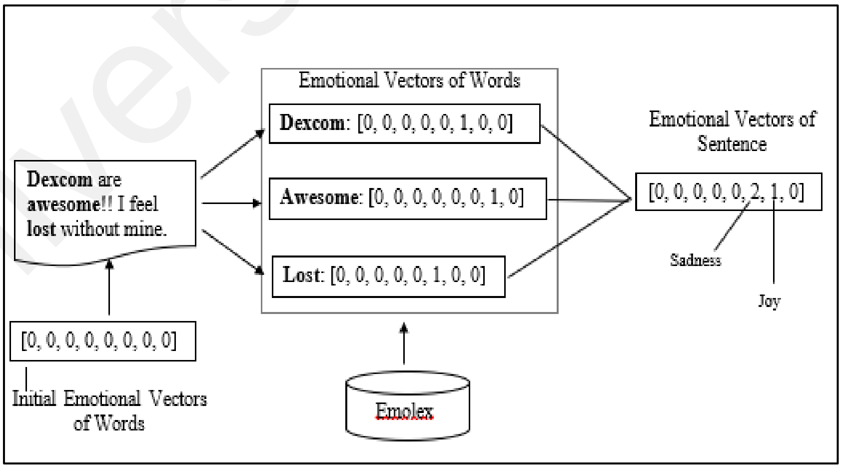


Figure 3.23 Graphical depiction of emotion vector assignment

As it can be observed from Figure 3.23, the emotion with the highest number assigned is *sadness* with a count of 2 and *joy* showing a count of 1 [0,0,0,0,0,**2**,1,0]. Emolex was created with the assumption that a sentence can have multiple emotions attached to it, therefore it is unable to assign a post to a dominant emotion. To overcome this issue, this research adopted the bootstrapping methodology using Word2Vec Continuous Bag of Words (CBOW) model. Annotated corpus from the dataset in this research was used to create a seed extension to produce a higher accuracy for emotion detection. The seed extension method measured the resemblance of lexicon words from Emolex and the annotated dataset using the CBOW model. When a similarity of more than 70% is achieved between the two, the sentence is annotated using the emotion detected from Emolex. Nonetheless, in the case of Emolex lexicon matching to one or more annotated data, the algorithm will select the annotated emotion whose similarity is higher and assign it as the dominant emotion (Figure 3.24).

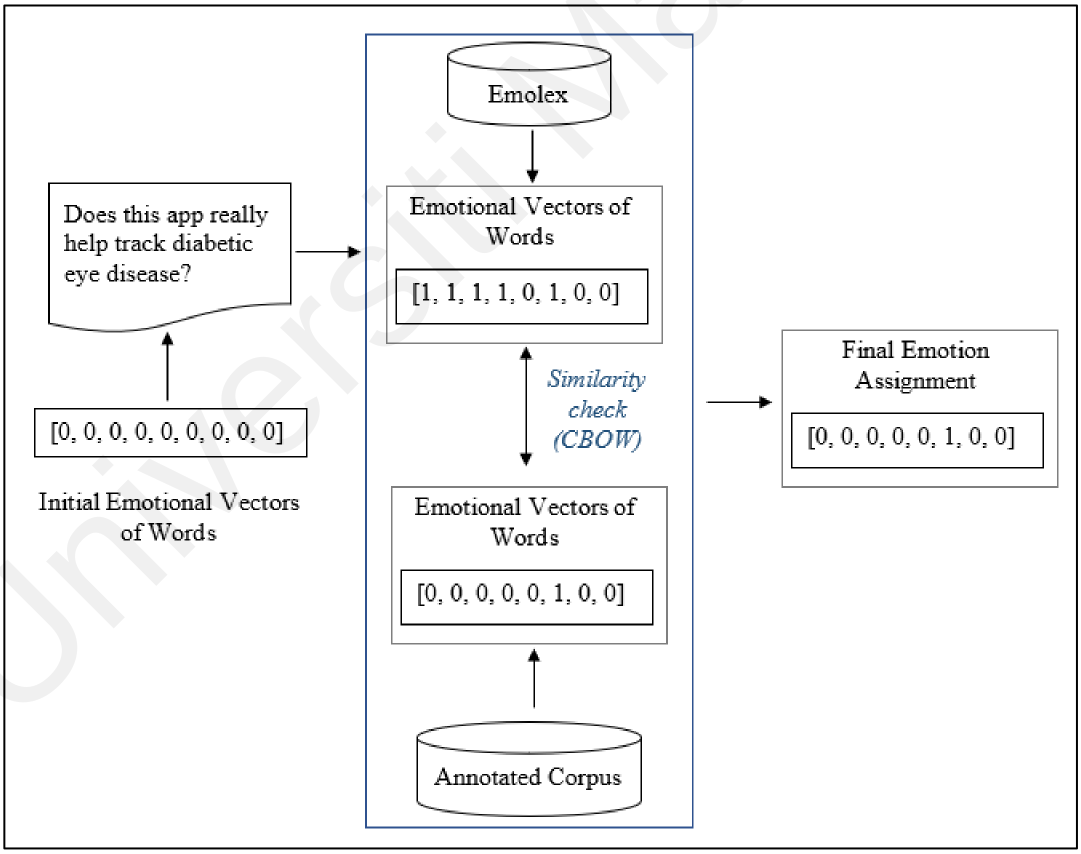


Figure 3.24 Emotion assignment

Once the similarity has been cross checked, a Sequential Minimal Optimization algorithm (SMO) and Support Vector Machine (SVM) supervised machine learning algorithms were adopted to train the classifier. Results for this tier will be discussed in the following chapter.

3.4.3.5 Summary of the Multi-Tier STEP Classification Framework

This subsection serves as a summary of how posts were classified within the proposed STEP framework. Figure 3.25 shows how a post moves through the tiers' framework.

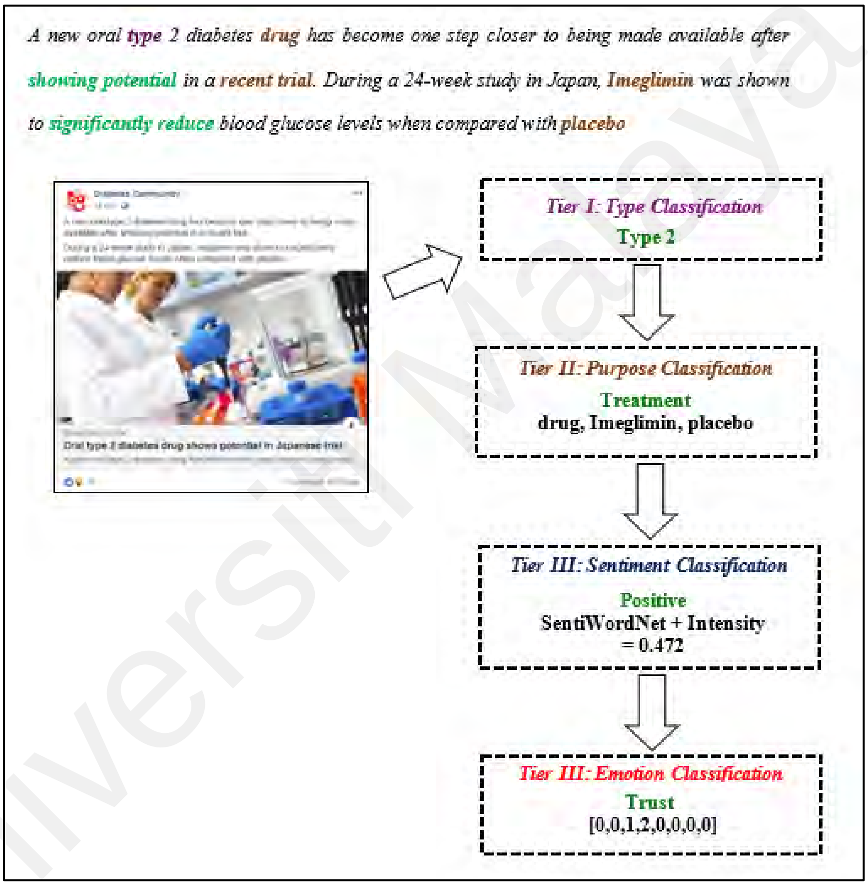


Figure 3.25 Post Classification through proposed STEP Framework

In the first tier of type classification, as the post contains the word "Type 2", it will automatically be classified as a type 2 post. Next is the the purpose classification. With respect to keywords like drug, Imeglimin and placebo, the algorithm would recognize these words as feature words belonging to the treatment label, hence classify this as Treatment. The sentiment is then determined using SentiWordNet and the assignment of weights with respect to the number of behaviors (i.e. 1 comment and 4 shares). Therefore, using Eq. 8 in Section 3.4.3.3, the sentiment polarity is calculated as 0.472. Finally, the emotion detected is Trust based on the highest count for the specific emotion (i.e. 2).

The following section describes the experimental setup and discusses the evaluation metrics of validating and verifying the results of this research.

3.5 Phase III: Evaluation Metrics

Two evaluation methods are used: standard evaluation metrics and comparing to benchmark models. The performance of each tier is done separately as each tier adopts a different technique. The performance was compared in terms of True Positive Rate (TPR), False Positive Rate (FPR), accuracy, F1-score and Area Under Curve (AUC). Two different confusion matrixes were used for binary classification of type classification (Table 3.8) and multi-label classification (Figure 3.26).

Table 3.8 Confusion Matrix for binary classification

		Predicted	
		False	True
Actual	False	TN	FP
	True	FN	TP

**TN = True Negative, TP = True Positive, FN = False Negative, FP = False Positive*

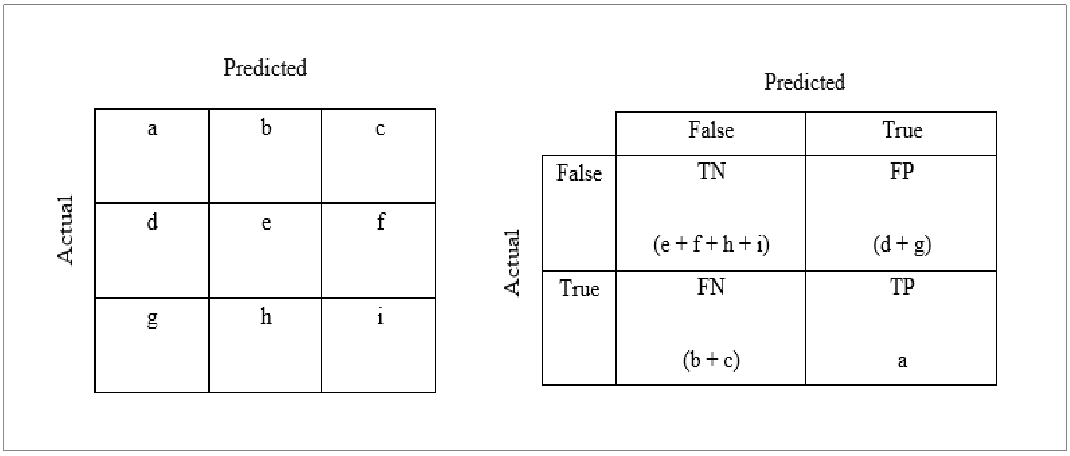


Figure 3.26 Multi-Label Confusion Matrix (Ruuska et al., 2018)

Table 3.8 and Figure 3.26 is generated from the following four measures:

- True Positive (TP) – Number of correctly classified data that belongs to class
- True Negative (TN) – Number of correctly classified data that do not belong to class
- False Positive (FP) – Number of incorrectly classified data as belonging to class
- False Negative (FN) – Incorrectly classified data that were not classified as class data

The evaluation was calculated using a ten-fold cross validation where data is divided into ten subsets and the holdout method is reiterated ten times. In every round, a single subset is taken as test set at a time while the balance nine subsets are merged to form the training set (Idrees, Rajarajan, Conti, Chen, & Rahulamathavan, 2017). Error encountered within all ten rounds are averaged out to produce the final output. This warrants each instance is included minimally once in the test set and nine times in the training set. Below are the equations for True Positive Rate (TPR), False Positive Rate (FPR), accuracy, F1-Score and Area Under Curve (AUC) respectively adopted from Idrees et al. (2017), D. Anand and Naorem (2016) Ruuska et al. (2018) and Bradley (1997)

$$TPR = \frac{TP}{TP + FN} \quad \text{Eq. 9}$$

$$FPR = \frac{TP + TN}{TP + FN + FP + TN} \quad \text{Eq. 10}$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad \text{Eq. 11}$$

$$F1 - Score = \frac{2\ x\ precision\ x\ recall}{precision + recall} \quad \text{Eq. 12}$$

$$AUC = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \qquad \text{Eq. 13}$$

F1-Score is used to calculate the precision-recall curve where the scope ranges between 0 (worst results) to 1 (best results). C.-H. Lee (2018) mentioned a higher F-measure score indicates a better quality of classification. Area Under Curve (AUC) has been recognized as an important evaluation metric when it comes to verifying a classification model's performance. It is said that the higher the AUC, the better the model is at predicting the correct classification (J. Chen, Chen, Wu, Hu, & Pan, 2017; Mountassir, Benbrahim, & Berrada, 2012). Similar to F-measure, the range of AUC is between 0 (worst performance) and 1 (best performance).

Apart from the above equations, different benchmark datasets have been used in order to validate each tier separately. The evaluation of each tier is done separately as each tier adopts a different technique. The following subsections will look into the experimental set up and evaluation for each of the tiers of the proposed STEP framework. All benchmark datasets below have been compared and the results of those comparisons will be discussed in the following chapter.

For ease of understanding, the naming convention set for benchmark models as well as proposed STEP classification framework for evaluation purpose is M_x where *M* represents Model and *x* represents the model's name. Each model name will be according to its author for example M_{Salas} would mean it's the benchmark model obtained from the author Salas-Zárate, Medina-Moreira, Lagos-Ortiz, Luna-Aveiga, Rodríguez-García, et al. (2017). For the case of the proposed STEP framework, the naming convention would be M_{STEP-x} where *x* represents the tier upon which the evaluation is being discussed. For example, M_{STEP-t} would mean the evaluation is for Type Classification of the STEP framework.

3.5.1 Tier I: Type Classification Evaluation

For this tier, the dataset of Reichert, Kristensen, Mukkamala, and Vatrapu (2017) and Salas-Zárate, Medina-Moreira, Lagos-Ortiz, Luna-Aveiga, Rodríguez-García, et al. (2017) were used to evaluate type classification. Reichert et al. (2017) used a supervised machine learning approach to analyze type 2 online health related forum scripts for both sentiment and emotion analysis. Salas-Zárate, Medina-Moreira, Lagos-Ortiz, Luna-Aveiga, Rodríguez-García, et al. (2017) conducted an aspect level type 1 diabetes diagnosis from tweets. Both the authors were kind enough to share their models, thus, both models will be used as benchmark for this tier of evaluation. Experiments within this tier were conducted using models as follow:

- $M_{Reichert}$: classification using Reichert et al. (2017) benchmark model

- M_{Salas}: classification using Salas-Zárate, Medina-Moreira, Lagos-Ortiz, Luna-Aveiga, Rodríguez-García, et al. (2017) benchmark model

- M_{STEP-t}: classification using proposed Multi-tier STEP classification framework

3.5.2 Tier II: Purpose Classification Evaluation

The evaluation of this phase used three multi-label performance measure (Read, Pfahringer, Holmes, & Frank, 2011) namely Hamming Loss, 0/1 Loss and accuracy. The Hamming Loss (Eq. 14) treats each label as a distinct binary evaluation while the 0/1 Loss (Eq. 15) measure states any predicted label must match true set of labels (c) exactly.

$$Hamming\ Loss = 1 - \frac{1}{NL}\sum_{i=1}^{N}\sum_{l=1}^{L} 1(c_l^i = \hat{c}_l^{\,i}) \qquad \text{Eq. 14}$$

$$0/1\ Loss = 1 - \frac{1}{NL}\sum_{i=1}^{N} 1(c^i = c^{\wedge i}) \qquad \text{Eq. 15}$$

The accuracy measure for multi-label classification (Eq. 16) was introduced by Godbole and Sarawagi (2004) and has been used as the standard evaluation technique since (Elghazel, Aussem, Gharroudi, & Saadaoui, 2016; A. U. R. Khan, Khan, & Khan, 2016; C.-H. Lee, 2018; S. M. Liu & Chen, 2015).

Eq. 16

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{c^{i} \wedge c^{\wedge i}}{c^{i} \vee c^{\wedge i}}$$

3.5.3 Tier III: Sentiment Classification Evaluation

The evaluation on this phase will compare baseline SentiWordNet score without the added intensity from Facebook behavior against with added intensity. Experiments within this tier were conducted using models as follow:

- M_{SWN}: Baseline without added intensity
- M_{STEP-s}: Baseline with added intensity of likes, shares, comment and reactions

3.5.4 Tier III: Emotion Classification Evaluation

Similar to sentiment classification evaluation, a comparison between baseline Emolex and proposed STEP framework was conducted. Experiments within this tier were conducted using models as follow:

- M_{EMO}: Baseline using Emolex
- M_{STEP-e}: Baseline with CBoW similarity checks.

3.6 Summary

This chapter described in detail the methodology adopted for this research. Each phase has been described in detail and each tier was looked into separately. Each phase of the methodology has been explained in detail form the pre-processing, the actual proposed STEP framework itself and the evaluation phase.

This chapter began with a description on the data source used for classification. This includes the pages from which the corpus was sourced and the amount of data that was crawled. The standard pre-processing of tokenizing, stemming and POS Tagging was discussed with example used from the data source itself. Out of the amount of data crawled, a large chunk of it had to be removed through the cleaning process. 6,000 were received back after human annotation which was then used as the labelled dataset for the purpose of testing and training the proposed STEP framework. Krippendorf alpha was calculated to determine the inter-coder agreement.

A topic modelling experiment was also explained in this chapter that helped to determine the number of classes that can be identified from the extracted corpus itself. The experiment generated a number of topics that provided the foundation on the number of classes for purpose classification.

The methodology applied in each phase of the proposed STEP framework was looked into closely. For type classification, the contribution was on the lexicon dictionary created that caters to all three types of diabetes. The next tier looked into purpose classification where a co-training Multinomial Naïve Bayes classification algorithm was adopted using weighted feature selection. The weighted information gain feature selection method allowed for weights to be redistributed for features that have been wrongly classified while the co-training looked to feed feature into label classes thus helping the feature to be correctly classified in the right label. The mathematical formula proposed for sentiment intensity was also discussed where samples of the calculation was also presented. For emotion classification, the common bag of words method was used to determine which emotion was dominant among all and thus classify the emotion accordingly. The last section looked into the experiment set up and evaluation used to evaluate the proposed STEP framework. Benchmark models for comparison as well as standard evaluation metrics such as F1-Score and AUC are going to be used commonly in the following chapter. The following chapter will take a closer look into the results using the evaluation techniques introduced in this chapter.

CHAPTER 4: RESULTS & DISCUSSION

This chapter presents the results of all the experiments conducted in this research. The results are displayed from preliminary to experiments carried out at each phase of the research. A brief overview on the results of the survey conducted amongst medical professionals for the purpose of annotated data collection and verification on purpose classification will also be provided in this section. As explained in section 3.5, the evaluation phase involved two evaluation metrics, namely F1-Score and area under curve (AUC). Comparison between benchmark model is also provided in this chapter. The results obtained within each tier of the proposed framework are discussed as well.

The rest of this chapter is organized as follows: results of the data pre-processing phase in identifying forms of pre-processing to be adopted are discussed, followed by the preliminary experiment results for topic modelling, and individual results for each of the three tiers of the proposed framework. This is then followed by comparisons of the evaluation metrics as well as comparison against the benchmark sets. Apart from that, analysis related to the diabetes dataset, especially those related to the user behaviors found on Facebook is also included in this chapter.

4.1 Data Preprocessing Results

Literature has stated the importance of data cleaning and pre-processing when conducting a sentiment analysis study. It is crucial to select the correct form of pre-processing techniques for the dataset used as too much or too little pre-processing may affect the outcome of the classification (Haddi et al., 2013; Singh & Kumari, 2016). Therefore, in order to identify the types of pre-processing that can be adopted for this research, a simple sentiment analysis experiment to check different forms of pre-processing was conducted using the diabetes corpus.

The experiment was conducted using Weka 3.8. Table 4.1 shows the pre-processing techniques applied at this stage of experiment. All the techniques applied were supported by built-in functions in Weka.

Table 4.1 Preprocessing Technique Applied

Pre-processing steps	Technique Applied
Tokenization	Unigram, bigram and n-gram (3, 4 and 5-gram)
Stemming	Snowball Stemmer
White space removal	White Space Tokenizer
Stop word and punctuation removal	Rainbow List
POS Tagging	Weka POS Tagger

Four commonly used classifiers; namely Naïve Bayes (NB), Support Vector Machine (SVM), k-Nearest Neighbor (KNN) and C4.5 Decision Tree (Krouska et al., 2016) and 1, 000 randomly selected posts from the corpus were used for the purpose of this experiment. The cross-validation was set to ten-fold whereas k = 17 was used as the optimal k value for KNN. For comparison reasons, two built-in feature selection mechanisms were used for this experiment, i.e. no filter (all attributes created using StringToWordVector) and InfoGainAttributeEval (IG). The difference of using a feature selection compared to no feature selection could be seen through the time taken to classify the posts whereby feature selection mechanisms helped to speed up the overall training process. Table 4.2 showcases the results of the experiment.

Table 4.2 Classifier Accuracy using Pre-processing Techniques

N-gram	Attribute	Classifier			
		NB	SVM	KNN	C4.5
Unigram	No Filter	75.21%	78.49%	71.41%	75.26%
	IG	**89.08%**	81.93%	72.20%	77.33%
Bigram	No Filter	87.70%	89.41%	79.91%	79.62%
	IG	**91.63%**	90.59%	84.40%	83.20%
Trigram	No Filter	90.93%	92.59%	87.21%	87.68%
	IG	93.42%	**95.08%**	88.92%	88.67%
4-gram	No Filter	62.34%	65.30%	58.12%	57.26%
	IG	65.11%	**67.71%**	61.60%	65.45%
5-gram	No Filter	50.21%	52.42%	49.96%	48.29%
	IG	54.13%	**55.63%**	51.78%	50.64%

From the table above, it can be seen that unigram, bigram and trigram outperformed the other two n-grams in terms of accuracy. This is in line with many literature that stated using one – three grams of n-gram processing produces more accurate classification, and as the number of n increases, the accuracy decreases (Gull, Shoaib, Rasheed, Abid, & Zahoor, 2016; Haddi et al., 2013; Singh & Kumari, 2016). The other pattern that can be observed is the use of feature selection which plays an important role in classifying text. The addition of feature selection not only helps the algorithm to run at a faster pace but also removes redundant attributes that would otherwise not contribute towards classification. Therefore, the usage of feature selection produced more accurate results in comparison to not using any.

Using this knowledge in hand, it became apparent that in the following experiment, the number of n-grams that would produce the best results would be between two and three with unigram coming in very close. Therefore n-gram were used for classification purpose for all three tiers. It was also found that including feature selections would be crucial to improve accuracy. The following section will discuss the results of the topic modelling experiment.

4.2 Topic Modelling Results

The top ten results are displayed as a sample together with the corresponding occurrences (Table 4.3). This term frequency table is the output of the document term matrix of the Latent Dirichlet Allocation (LDA) model where frequency refers to the number of times that particular term has been found within the corpus.

Table 4.3 Term Frequency

No	Term	Frequency
1	Dexcom	2, 480
2	Metformin	2, 120
3	Insulin	1, 988
4	Lethargic	1, 846
5	EpiPen	1, 726
6	Keto	1, 522
7	DKA (diabetic ketoacidosis)	1, 197
8	Worried	986
9	Upsetting	740
10	Clinical	699

The topics produced within this stage provided the basis for the number of classes for purpose classification. The next section looks into the results churned out within each tier of the proposed sentiment-type-emotion-purpose (STEP) framework.

4.3 Sentiment-Type-Emotion-Purpose (STEP) Framework Results

Each sub-section below will look into the individual tiers of the proposed classifier and discuss the results separately as each tier adopts a different methodology as discussed in Chapter 3.

As mentioned in Chapter 3, the naming convention set for benchmark models as well as proposed STEP classification framework for evaluation purpose is M_x where M represents Model and x represents the model's (author) name. For the case of the proposed STEP framework, the naming convention would be M_{STEP-x} where x represents the tier upon which the evaluation is being discussed. For example, M_{STEP-t} would mean the evaluation is for Type Classification of the STEP framework.

4.3.1 Tier I: Type Classification

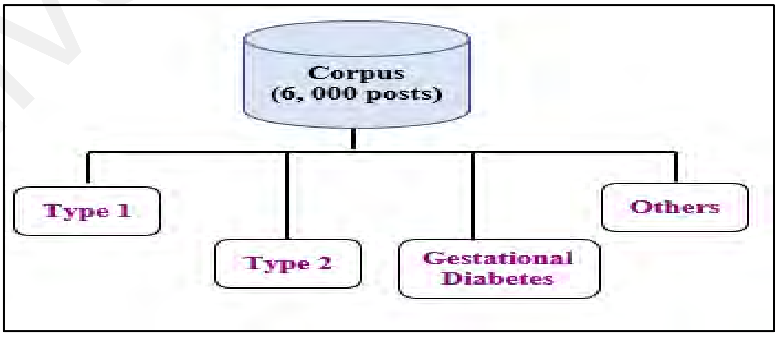


Figure 4.1 Tier 1: Type Classification

Figure 4.1 is a depiction of the first tier of the proposed STEP framework. This section will discuss the results achieved within this tier of classification. An experiment was conducted to learn if the combination of any two forms of n-gram would increase the F1-

Score of the classification. Therefore, several different combinations were tested (Table 4.4) using four of the most widely used classifiers in text classification (Allahyari et al., 2017) namely Naïve Bayes, Support Vector Machine, k-Nearest Neighbor and Logistic Regression.

Table 4.4 Classification F1-Score Using Different N-grams

N-gram	Classifiers			
	NB	SVM	KNN	LR
Unigram	0.62	0.63	0.63	0.60
Bigram	0.67	0.62	0.66	0.62
Trigram	0.56	0.54	0.50	0.53
Unigram + Bigram	**0.72**	**0.68**	**0.70**	**0.65**
Bigram + Trigram	0.63	0.65	0.66	0.61
Trigram + Unigram	0.55	0.58	0.58	0.52

**NB = Naïve Bayes, SVM = Support Vector Machine, KNN, K-Nearest Neighbor, LR = Logistic Regression*

The results in Table 4.4 revealed that the combination of unigrams and bigrams produced the best results compared to other combinations. Past literature has also agreed on the use of unigram and bigram to produce more accurate classification results (Tripathy et al., 2016). This is due to tokenization and stemming during the pre-processing phase which increases the probability of the combination to appear in the training dataset compared to any other combinations (Aisopos, Tzannetos, Violos, & Varvarigou, 2016; Tripathy et al., 2016; J. Vilares, Vilares, Alonso, & Oakes, 2016). As supported by literature, Naïve Bayes classifier applies an index on every word position of a text document which allows the algorithm to recognize factors such as means and variance of variables that are crucial for accurate text classification (Bilal, Israr, Shahid,

& Khan, 2016; Diab & El Hindi, 2017; A. U. R. Khan et al., 2016).Therefore, in cases where the training dataset is limited, Naïve Bayes has the ability to produce a higher accuracy compared to other machine learning algorithms (Kadhim, 2019). With respect to the results obtained (Table 4.4), this research adopted the Naïve Bayes classifier along with unigram and bigram feature selections.

Another set of experiments were done using the benchmark datasets ($M_{Reichert}$ and M_{Salas}). As explained in Chapter 3, the comparison was only conducted for Type 1 and Type 2 classification as M_{Salas} comprises of Type 1 tweets while $M_{Reichert}$ contains Type 2 online health forum dataset. Based on the results observed in Table 4.5, the classification results for Type 2 ($M_{Reichert}$) proved to be more encouraging compared to Type 1 (M_{Salas}). When analyzing the online health forum dataset, it was discovered that they were almost like Facebook posts in terms of length and jargon used within the text itself. Therefore, the classification results obtained using $M_{Reichert}$ dataset is almost similar to M_{SEP-t} (T2) results. Tweets on the other hand, require a different form of data cleaning and pre-processing (Chandra Pandey, Singh Rajpoot, & Saraswat, 2017; Tellez et al., 2017) which affected the results obtained.

Table 4.5 Type Classification Comparison Results

Dataset	Evaluation Metrics		
	F1-Score	Accuracy	AUC
M_{Salas}	0.48	0.53	0.53
$M_{Reichert}$	0.70	0.69	0.70
M_{SEP-t} (T1)	**0.77**	**0.76**	**0.77**
M_{SEP-t} (T2)	0.69	0.69	0.69
M_{SEP-t} (T3)	0.76	0.75	0.76

M_{Salas} = Type 1 tweets,
$M_{Reichert}$= Type 2 online health dataset,
M_{SEP-t} (T1)= Type 1 Proposed STEP Classifier dataset,
M_{SEP-t} (T2)= Type 2 Proposed STEP Classifier dataset,
M_{SEP-t} (T3) = Gestational Diabetes Proposed STEP Classifier dataset

When comparing the classification results of each individual type of diabetes (Table 4.5), it shows that the proposed framework was able to classify Type 1 more accurately in comparison to Type 1 and Type 3. This may be contributed to the lexicon used within this tier which is an extension from the Type 1 diabetes ontology used by El-Sappagh and Ali (2016). The lexicon dictionary contains more words that cater for Type 1 diabetes, and thus the ability to match more words for Type 1 diabetes may have improved the classification for this type. Gestational diabetes (i.e. Type 3) classification scores also proved to be better compared to Type 2. Again, this is probably due to the keyword matching between the lexicon dictionary and the words used within the dataset.

Figure 4.2 is a graphical depiction of the number of posts classified per type. Posts that could not be classified into any of the three types of diabetes were categorized as Other (N = 1, 111). From the figure it can be observed that the majority of the posts belong to Type 1 followed by gestational diabetes (Type 3) and Type 2. The next section will look into the results of purpose classification where the number of data that is carried from this tier into the next will be 4, 889 posts.

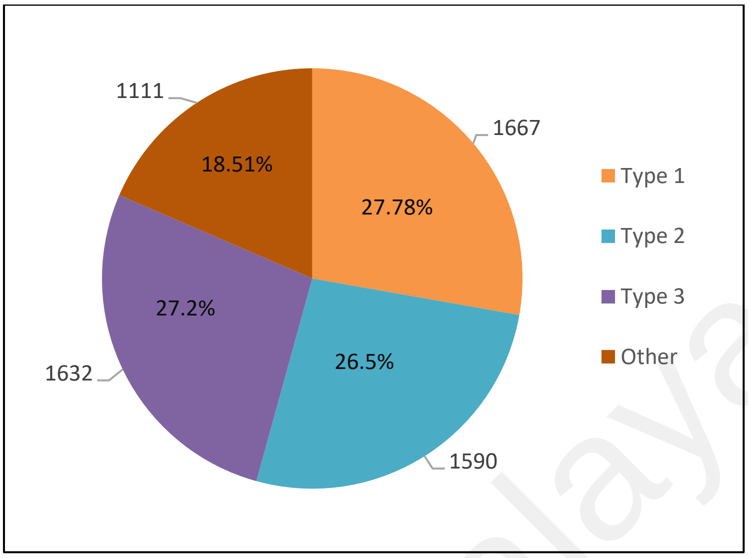
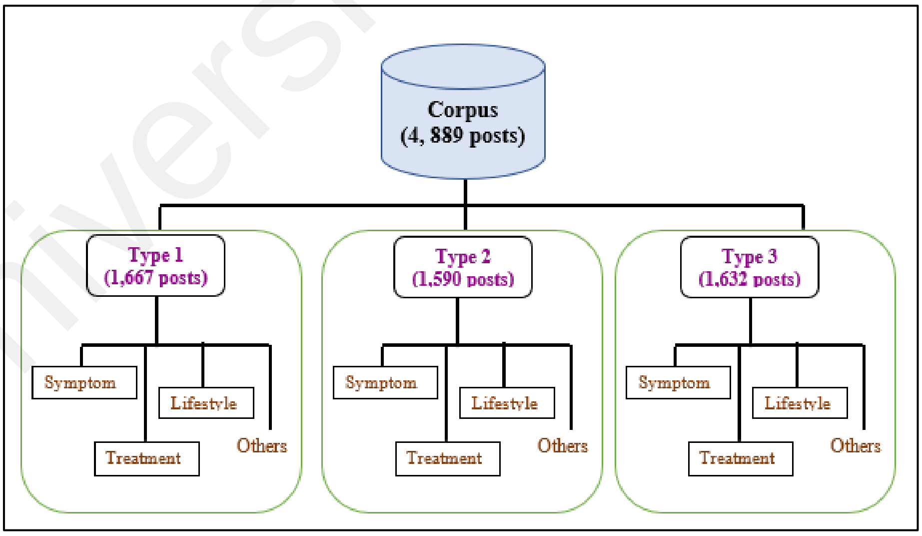


Figure 4.2 Posts Classified Per Diabetes Type

4.3.2 Tier II: Purpose Classification



**Type 3 = Gestational Diabetes*

Figure 4.3 Tier II: Purpose Classification

Figure 4.3 shows the second tier of the proposed STEP framework. After the classification in tier 1, the number of posts available for classification within this post were 4,889. Data that have been classified as other within the tier above (type classification) were not brought forward to this tier. This tier will classify post according to Symptom, Treatment and Lifestyle. Similar to the above tier, posts that were not able to be classified between the aforementioned classes were classified as Other.

The literature defines features as terms that appear within a textual dataset and frequency as the number of times a term appears (Akhtar et al., 2017; Rehman, Javed, & Babri, 2017). Therefore, a feature selection experiment was conducted to determine the type of feature selection that would suit best for purpose classification and the number of features that would produce the most accurate results. Six of the most widely used feature selection techniques (Rehman et al., 2017); namely Odds Ratio (OR), Information Gain (IG), Chi Square (CH), Distinguishing Feature Selector (DFS), Gini Index (GINI), Poisson Ratio (POIS) were compared using three of the most widely used classifiers in text classification (Allahyari et al., 2017) namely Naïve Bayes, Support Vector Machine and Logistic Regression.. Results of the experiments are as displayed for F1-Scores against number of features using different classifiers (Figure 4.4, Figure 4.5, Figure 4.6) while Table 4.6 depicts the best feature selection technique based on the number of features (per hundred).

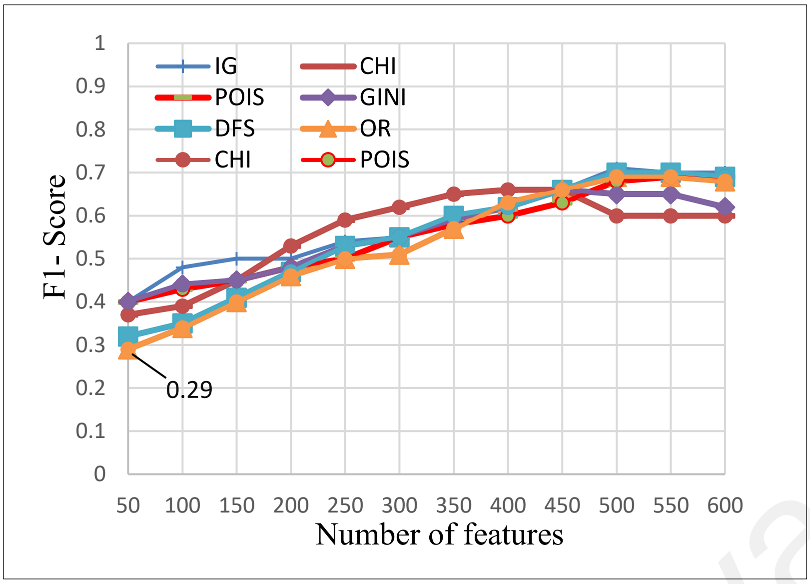


Figure 4.4 F1 Score Using Naïve Bayes

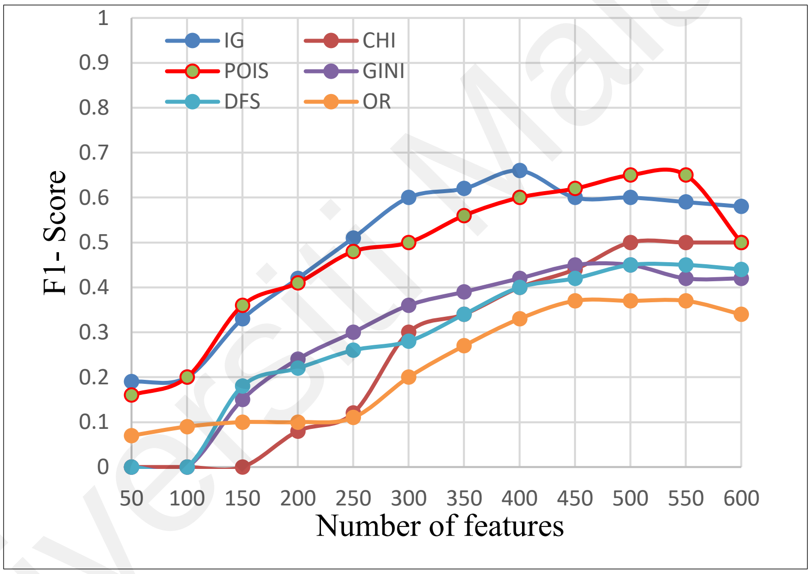


Figure 4.5 F1 Score Using Support Vector Machine

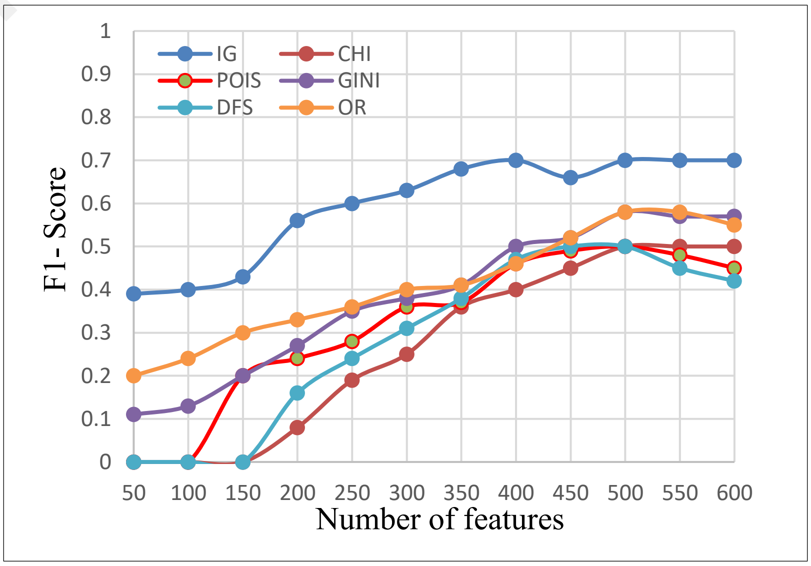


Figure 4.6 F1 Score Using Logistic Regression

Table 4.6 Feature selection technique producing highest F1 Score

Classifier	Features (per hundred)					
	100	200	300	400	500	600
Naïve Bayes	IG	CH	CH	CH	IG	IG
Support Vector Machine	IG	IG	IG	IG	POIS	IG
Logistic Regression	IG	IG	IG	IG	IG	IG

**IG = Information Gain, CH = Chi Square, POIS = Poisson Ratio*

As it can be observed from the results shown in Figures 4.2, 4.3 and 4.4, the optimum F1 Score is achieved when the number of features were 500. Furthermore, in the trials conducted above, it was also noticed that amongst the three classifiers used, the results of Naïve Bayes (Figure 4.4) showed to be more promising. For Support Vector Machine (Figure 4.5) and Logistic Regression (Figure 4.6), the classifiers produced a zero F1 Score for features between 50 and 150. However, Naïve Bayes worst F1 Score for a feature selection technique tested was 0.29. This goes to show Naïve Bayes has the ability to identify features even if the number of features is set to as low as 50. Literature has also supported the fact that Naïve Bayes works best when it comes to multi-label or multi-class classifications (A. U. R. Khan et al., 2016; C.-H. Lee, 2018). Therefore, for this tier of classification, this research adopted Information Gain feature selection technique and Naïve Bayes classification algorithm.

From the topic modelling results discussed in section 4.2, the number of labels identified for purpose classification were initially ten (cause, symptom, exercise, modern treatment, traditional treatment, emotional support, financial aid, advice and promotion), However, with experiments conducted, it became evident that some labels were over lapping each other and caused the F1-Score of the classification to suffer. Therefore, to counter this problem, some labels had to be combined. Results of the trial experiments is as shown in Figure 4.7 below.

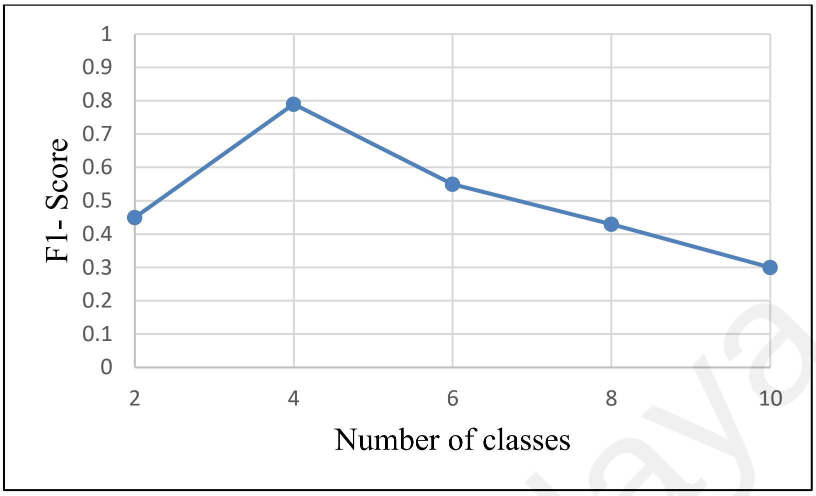


Figure 4.7 F1 Score with respect to number of classes

With reference to Figure 4.7, as the number of labels increased, the F1-Score decreased indicating the performance of the classifier was declining. This is because there was not enough labelled data for training. Therefore, efforts were made to combine similar label under a more general label which eventually improved the F1-Score distinctly. For example, instead of classifying *Metformin* as modern medicine and *herbal mixture* as traditional medicine, both (Metformin and herbal mixture) were instead classified into the label Treatment.

Once the labels were reduced to four (lifestyle, treatment, symptom and other), it was then discovered the classification within the labels leaned towards multi-label classification as discussed in Chapter 3. For example,

Herbal tea *first thing in the morning helps keep my blood sugar levels steady till I have breakfast.*

From the aforementioned sample, *herbal tea* can be labelled as treatment as well as lifestyle changes, thus rendering it as a multi-label problem. To overcome this, literature suggests the usage of Multinomial Naïve Bayes (MNB) (Aldoğan & Yaslan, 2017; C.-H. Lee, 2018; J. Lee & Kim, 2017). However, an experiment using MNB did not show promising FI-Scores and AUC as shown in Table 4.7. Hence, co-training with weighted Information Gain feature selection and string vectors were introduced, with improved results.

Table 4.7 F1-Score for Purpose Classification

	F1-Score			AUC		
	Symptom	Life Style	Treatment	Symptom	Life Style	Treatment
MNB	0.38	0.45	0.40	0.38	0.40	0.40
MNB + Co-Training + Weighted IG	0.48	0.51	0.45	0.47	0.51	0.42
MNB + Co-Training + Weighted IG + String Vectors	**0.61**	**0.57**	**0.57**	**0.58**	**0.58**	**0.47**

**MNB = Multinomial Naïve Bayes, IG = Information Gain*

The co-training algorithm works by identifying features as well as labels, meaning it was able to recognize terms such as "*Metformin daily*" as a bigram feature and classify it under *Treatment* label. Although the highest number of data available for purpose classification belonged to the *Treatment* label (Figure 4.9), the *Symptom* label produced the highest F1-Score and AUC. When analyzing the reason behind this difference, it was discovered *Symptom* had the most clearly defined boundaries in distinguishing it from the rest which allowed the algorithm to classify it easily. For example:

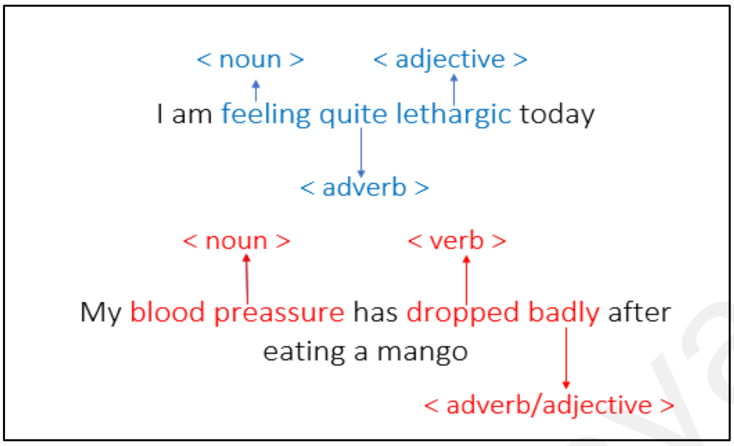


Figure 4.8 Sample text

The classifier picked on adjectives that either preceded or succeeded adverbs while cross checking for referenced nouns (Figure 4.8) within the post before classifying it under *Symptom*. Furthermore, the weighted information gain feature selection technique readjusted weights for wrongly classified features and because the features got more distinct for *Symptom* label compared to the other two, the F1-Score and AUC for symptom produced better scores. Similarly, identifying a post that distinctly belongs to *Lifestyle* and *Treatment* option was a little trickier as there were many posts that could either belong to one or the other. For example:

Having coffee after meals have helped me keep my blood sugar levels stable.

The above shows it is a lifestyle change to a diet but it can also be considered as a home remedy treatment. In order to further ascertain if there was a possible pattern that could help the algorithm distinguish between *Treatment* and *Lifestyle* label, 100 conflicted posts were sent for annotation. The Krippendorff alpha was determined at 0.62 which is not accepted as a proper agreement rate between the annotators (Krippendorff, 2011). However, if the amount of data available for training the algorithm was expanded, perhaps the F1-Score and AUC could be improved (Uysal, 2016).

As discussed in Chapter 3, the Hamming Loss, 0/1 Loss and Accuracy have been accepted as the standard evaluation metrics in evaluating multi-label classification algorithms (C.-H. Lee, 2018; Jun Li et al., 2016). Hamming Loss looks into the individual labels that have been incorrectly predicted, while 0/1 Loss looks at the set of labels in its entirety. Therefore, the whole set of labels in a sample post will be considered incorrect if it does not match the true set of labels. Table 4.9 depicts the results of each individual label based on these metrics.

Table 4.8 Other Evaluation Metrics for Each Label

	Symptom	Life Style	Treatment
Hamming Loss	0.087	0.233	0.316
0/1 Loss	**0.886**	**0.719**	**0.774**
Accuracy	0.701	0.681	0.683

The co-training algorithm used within this tier of classification used individual features as input into dependency label classifier. Therefore, the results of Hamming Loss are recorded as the lowest as it is calculated for label-based evaluation. The 0/1 Loss metric on the other hand, is designed for label-based evaluation and the co-training algorithm manipulates whole labels for classification purpose, therefore the results for the 0/1 Loss is recorded as the highest amongst the three.

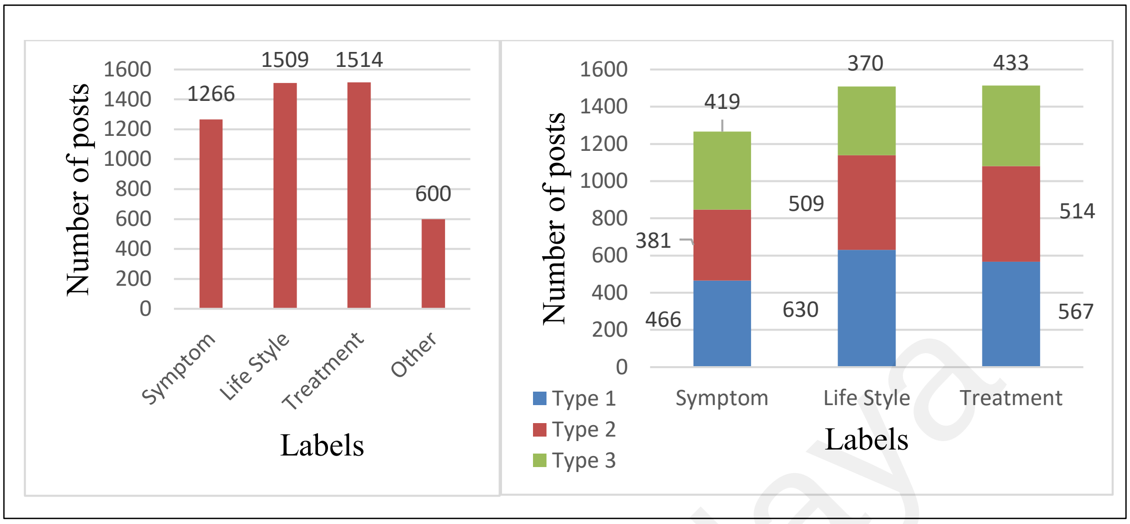


Figure 4.9 Posts Classified by Purpose

Figure 4.9 is a graphical depiction on the breakdown of data within the purpose labels. It can be observed that the purpose classifier was able to classify *Type 1* posts for *Life Style* label the most. This was an easy pick as the training data contained a lot of recipes and exercise options that were classified as *Life Style*. With the help of the weighted information gain feature selection, the algorithm kept reassigning weights to the labels that were incorrectly classified, hence improving its accuracy. The label *Symptom* has the least amount of data as the nature of the dataset contained less posts with respect to symptoms but more towards treatment options and changes that can be made to daily life that could help improve patients' quality of life.

The final analysis of this tier was based on the classifier's performance with respect to each type. Table 4.10 showcases the F1-Score obtained for each purpose (label) per diabetes type. *Life Style* for Type 2 diabetes scored the highest score amongst all while the lowest F1-Score was recorded by *Treatment,* also for Type 2. This was due to the nature of the training data that was fed into the algorithm. Since Type 2 diabetes is a much more controlled form of diabetes, hence the advice that comes from this column is much more related to healthy snack and diet options followed by home remedies that could help delay the consequences of the disease. The most prominent treatment known for Type 2 is Metformin, however, from the text itself, it is difficult to gauge if a treatment suggested is meant for Type 1 or Type 2, which explains the low F1-Score for Treatment of Type 2. The following sub-section will discuss results of the sentiment classifications.

Table 4.9 F1-Score for each Purpose label per Type

Type of Diabetes	F1-Score		
	Symptom	Life Style	Treatment
Type 1	0.58	0.70	0.70
Type 2	0.60	**0.71**	0.50
Type 3 (Gestational Diabetes)	0.63	0.63	0.56

4.3.3 Tier III: Sentiment Classification Results

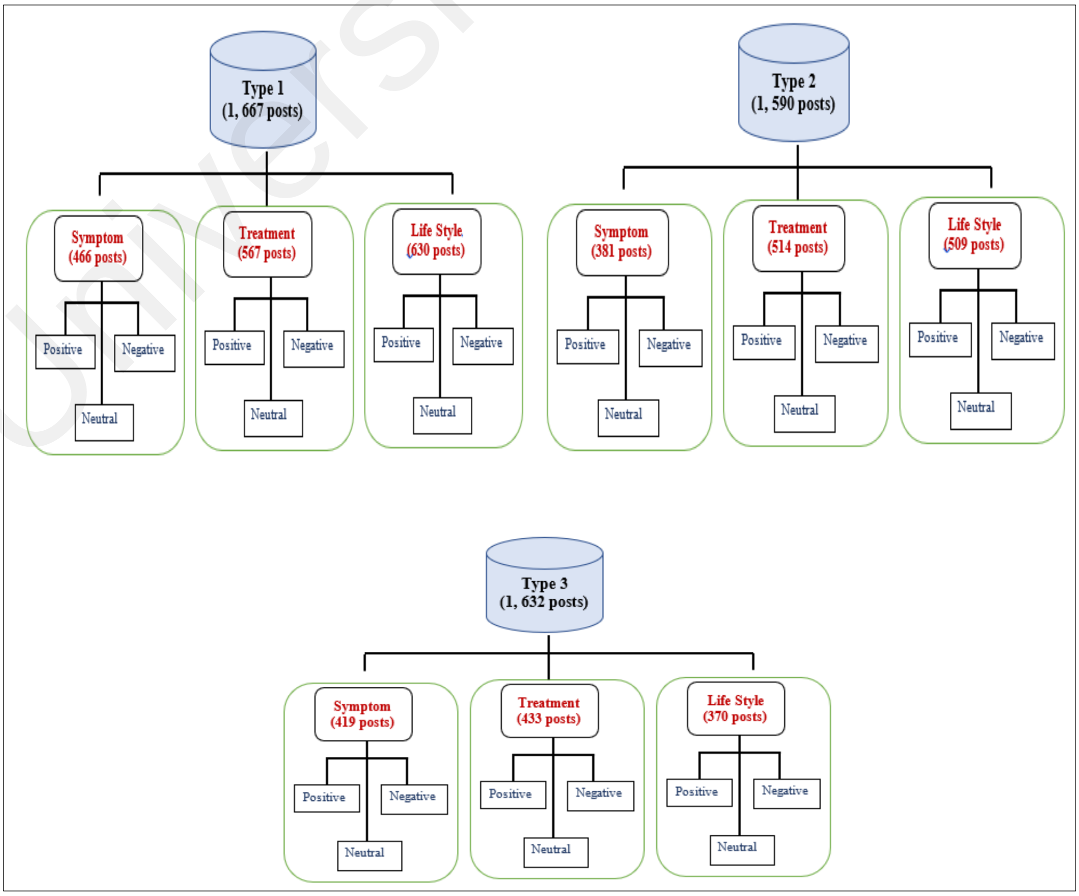


Figure 4.10 Tier III: Sentiment Classification

Figure 4.10 shows the third tier of the proposed STEP framework. After the classification in tier 2 (Purpose Classification), the number of posts available for classification for this tier can be seen in the figure (Figure 4.10). Data that have been classified as *other* within the tier above (purpose classification) were not brought forward to this tier. This tier will classify post according to Positive, Negative and Neutral.

Figure 4.11 shows the overall breakdown of total labelled diabetes data with respect to sentiment while Figure 4.12 looks into number of posts classified by sentiment according to individual labels (symptom, lifestyle and treatment) broken down by type (type 1, type 2 and gestational diabetes)

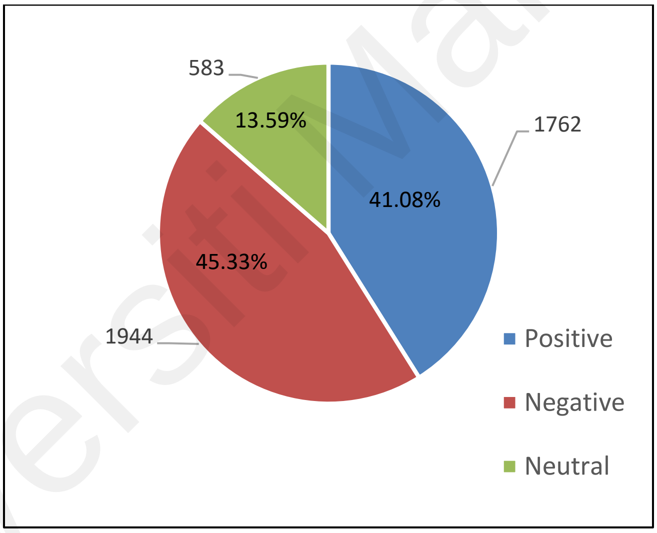


Figure 4.11 Total of labeled data for sentiment

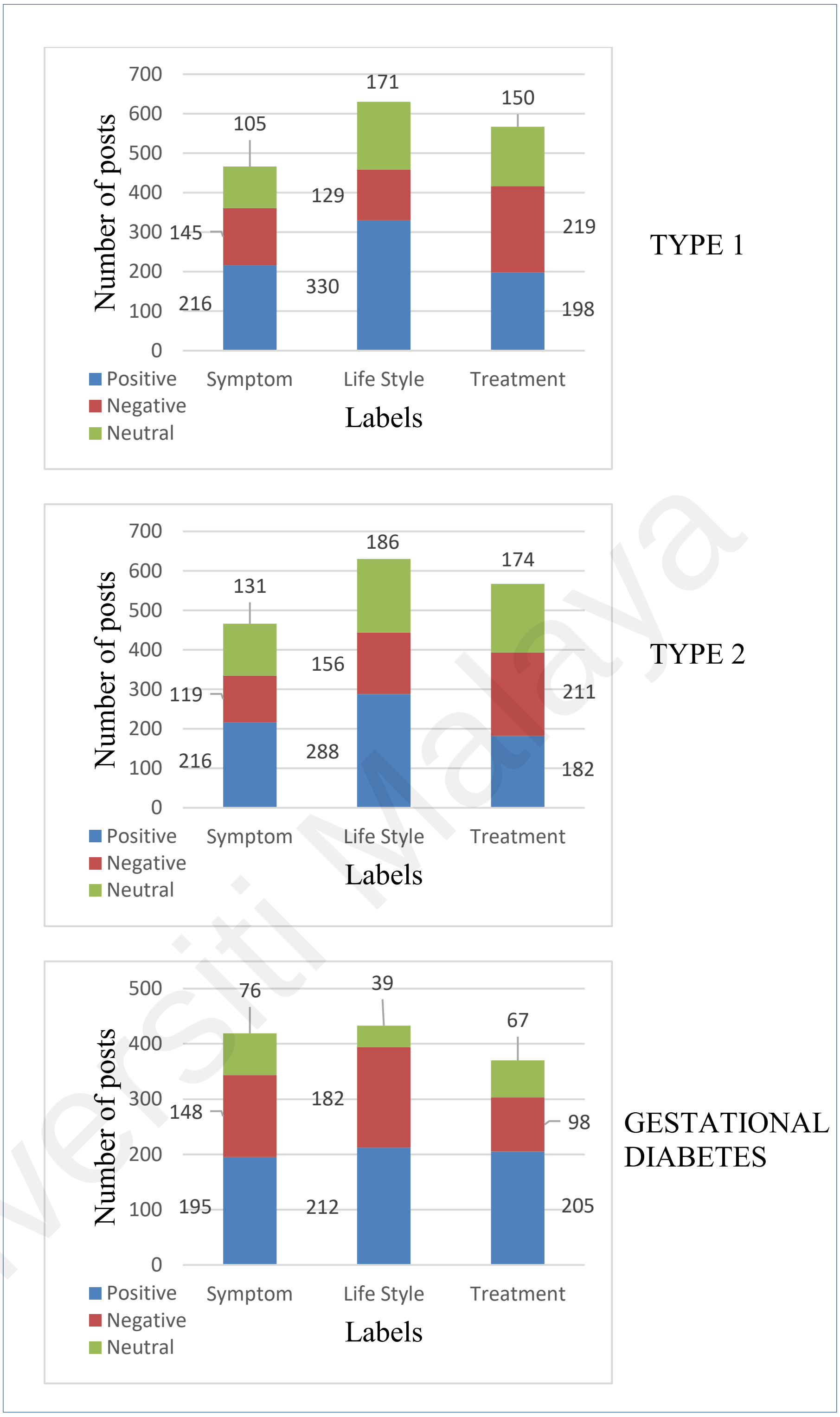


Figure 4.12 Number of post classification per Purpose classification per Type

Comparing Figure 4.11 and 4.12, it can be observed the most significant sentiment is the negative sentiment, however, when looked into each type, the highest number of negative posts come from Type 2 classification for the label treatment. This is due to the words used within the dataset that match the negative lexicons better. When the data was studied on the types of post (for type 2, treatment purpose), it was found users were particularly unhappy with the Metformin medication. This is due to the side effects that came with the medication and how not everyone responds to the medication in controlling their blood sugar levels. The least amount of neutral text belongs to gestational diabetes, this is because the posts within this type were very lengthy and detailed in description therefore the algorithm was able to pick out more words for classification using SentiWordNet and the usage of Facebook behaviors (*like*, *share*, *comment* and *reaction*). were used more within this type compared to the rest.

As explained in Chapter 3, one of the contributions of this research is in the form of a mathematical equation to calculate sentiment intensity using Facebook behaviors (*like*, *share*, *comment* and *reaction*). Therefore, in this sub-section the results were compared between the benchmark model (M_{SWN}) with no intensity calculation and with sentiment intensity (M_{STEP-s}). Tables 4.11 and 4.12 showcase the comparison results between the two aforementioned.

Table 4.10 Comparison results between M_{SWN} and M_{STEP-s}

	Metrics		
	F1-Score	AUC	Accuracy
M_{SWN}	0.68	0.64	0.64
M_{STEP-s}	**0.82**	**0.76**	**0.77**

**M_{SWN} = Baseline (SentiWordNet) without intensity,*
M_{STEP-s}= Proposed framework with intensity

With respect to Table 4.11, the proposed STEP framework with intensity (M_{STEP-s}) was able to better classify posts according to their sentiments compared to the benchmark (T_{SWN}) as the score improved from 0.64 to 0.77. This is contributed by the inclusion of intensity calculation that comes in the form of shares, likes, comments and reactions, indicating that each of these behaviors indirectly contributes to sentiment as supported in the literature as well (C. Kim & Yang, 2017; Quesenberry & Coolsen, 2018; Zell & Moeller, 2018).

Based on Table 4.12, the highest scores were achieved when classifying negative posts. When analyzing the data, it was found users tend to use stronger negative verbs when posting. The stronger negative verb carries a stronger negative score in SentiWordNet which is the baseline of the sentiment classifier proposed in STEP framework. Additionally, posts and comments that appeared to have stronger negative emotions evoked more people to comment and although the counter comments was laced with positivity, there was still hint of negativity within those counter comments which increases the number of negative comments compared to positive.

Table 4.11 Individual Experiment Breakdown

Metrics	M_{STEP-s}			M_{SWN}		
	Positive	Negative	Neutral	Positive	Negative	Neutral
F1-Score	0.66	**0.72**	0.63	0.60	0.51	0.55
Accuracy	0.65	**0.74**	0.69	0.62	0.55	0.61
AUC	0.65	**0.74**	0.68	0.62	0.55	0.61

*M_{SWN} = *Baseline (SentiWordNet) without intensity,*
M_{STEP-s}= *Proposed framework with intensity*

As the proposed framework is hierarchical, this tier also analyzed each purpose with respect to sentiment. The following results discussed will be related to the aforementioned.

Apart from comparing benchmark results for sentiment classification, an analysis on the posts of the corpus were also conducted. This was done using 4, 289 posts that were used for sentiment classification out of which 66% of the data were actually comments in reply of posts. Literature has already mentioned that social media users who log on to such platforms for health related purposes do so in order to feel connected and share experiences with one another (Fergie et al., 2016; Greene et al., 2011). Based on our data, the reaction buttons were used 51% of the time with the most used reaction being *love,* followed by *wow*, *sad*, *angry* and finally *haha*. This could be due to the age factor and nature of the group, so people are a little more sensitive and resort to a more cognitive approach to the usage of such reactions as compared to other social related groups. It was also found that positive related reaction buttons (*love* and *haha*) were more widely used compared to negative reaction buttons (*sad* and *angry*) as literature also indicates users on health related groups on Facebook are a lot more supportive of one another, hence the vibe around the groups tend to be more positive (Frison & Eggermont, 2015; Oh, Lauckner, Boehmer, Fewins-Bliss, & Li, 2013). Figure 4.13 depicts the breakdown on the number of posts, comments and reactions collected.

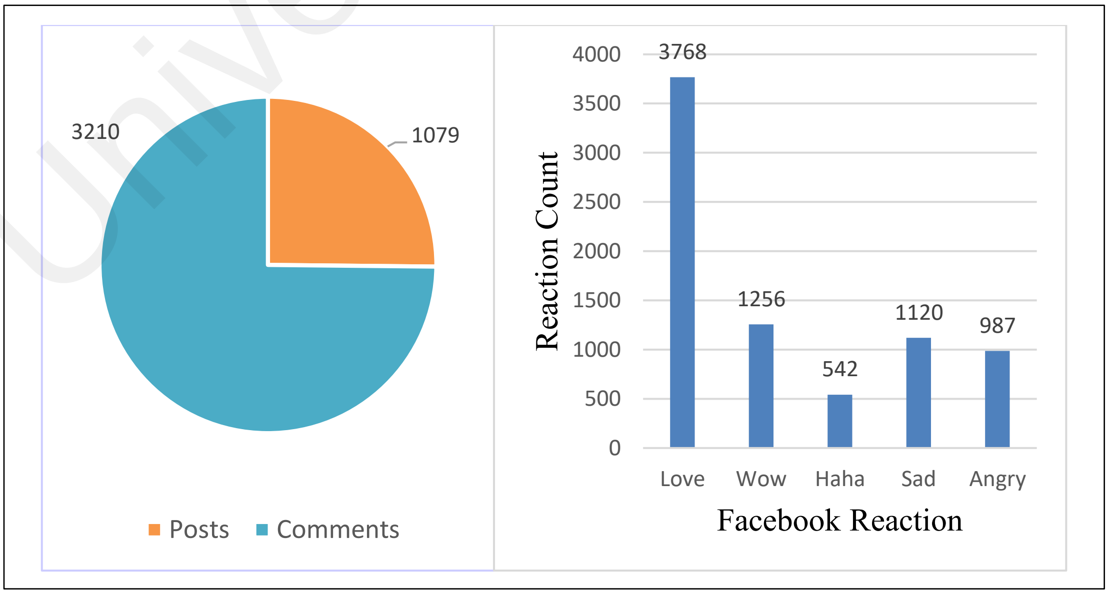


Figure 4.13 Number of posts, comments and reactions collected

As shown in Figure 4.14 below, users of the Facebook diabetes community were more likely to *share* a post compared to *like* a comment. Sharing requires a deeper cognitive thinking process, hence when it comes to information that users deem useful, the tendency of sharing that bit of information triggers the need to share (H. J. Kaur & Kumar, 2015; C. Kim & Yang, 2017; Quesenberry & Coolsen, 2018). Furthermore, each of the health-related groups on Facebook are monitored by a group of moderators, hence the trust on the items being posted is validated and users are more prone to share such information without hesitation.

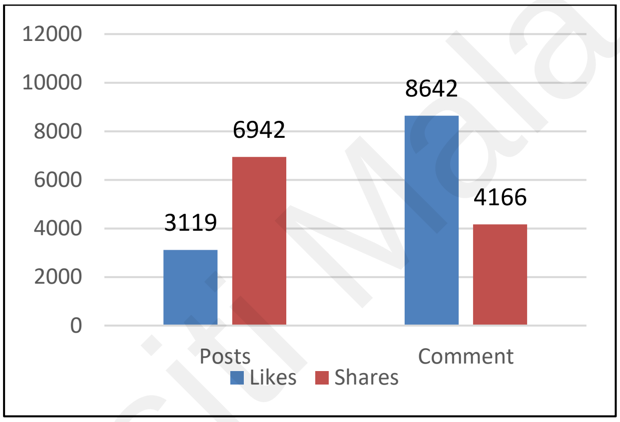


Figure 4.14 Number of *like* and *share* for posts and comments

Among other things, an interesting insight into the data analysis revealed the length of posts influences the number of *share* it garners (Figure 4.15). The inference is that a long post forces a user to spend time reading through it before deciding it worthy of being shared. In general, between all three *shares*, *comment* and *like*, there seems to be a threshold to the number of characters as to when users decide a post is not of interest to them to either be shared, commented or liked upon. When it comes to *like*, the shorter a post, the more *likes* it accumulates, however users are less likely to *comment* if a post is too long but are good to *share* it with others.

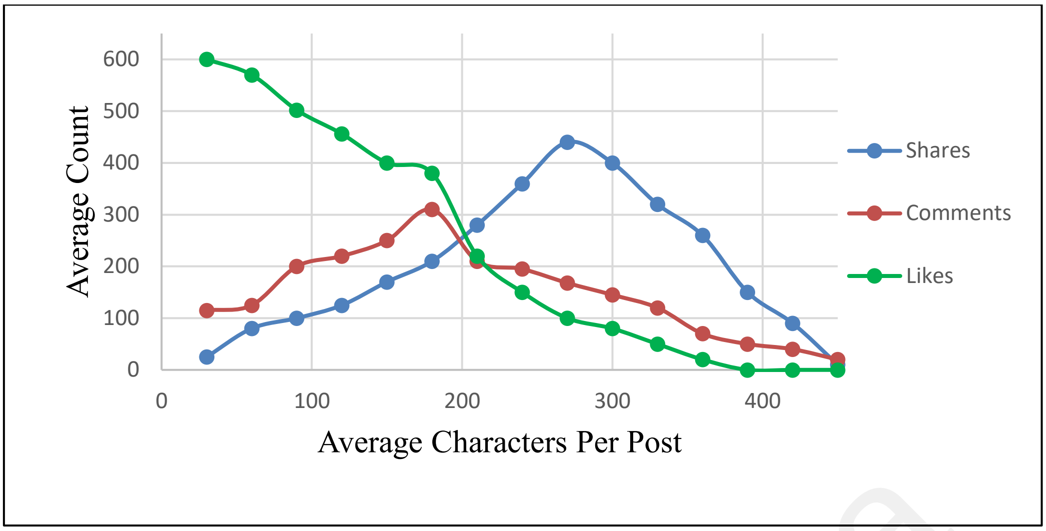


Figure 4.15 Number of shares with respect to number of characters in posts

The following sub-section looks into the final tier of classification which is the emotion classification.

4.3.4 Tier III: Emotion Classification Results

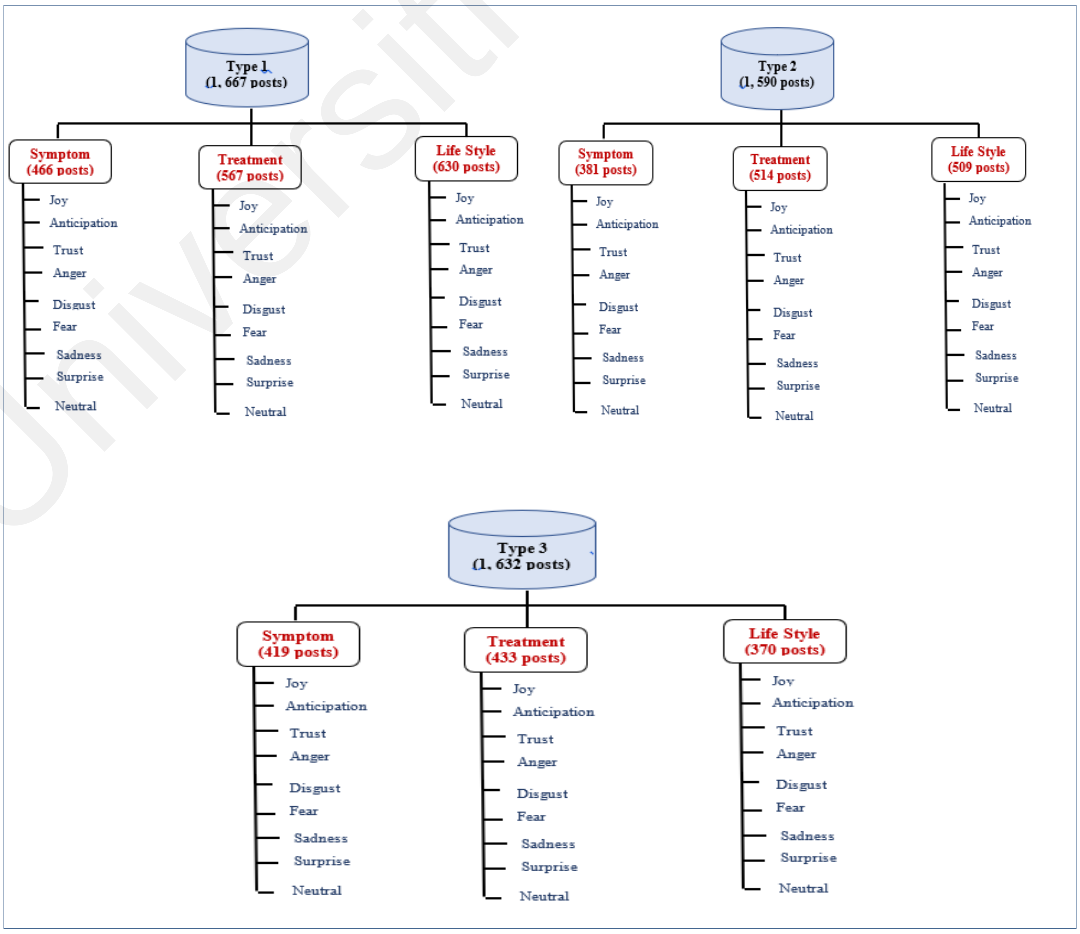


Figure 4.16 Tier III: Emotion Classification

Figure 4.16 shows the third tier of the proposed STEP framework. In the third tier, posts are classified according to both sentiment and emotion. These two elements (sentiment and emotion) lie within the same tier (Tier III). In the past, literature has treated sentiment and emotion as mutually inclusive instead of being two separate entities (Rosso et al., 2016; Rout et al., 2018). Therefore, for this stage of classification, the STEP framework classified both sentiment and emotion within the same tier. Posts that could not identify any emotion were labelled as Neutral. Figure 4.17 shows the general breakdown of the number of labelled data used for training and testing purpose within this tier for emotion classification.

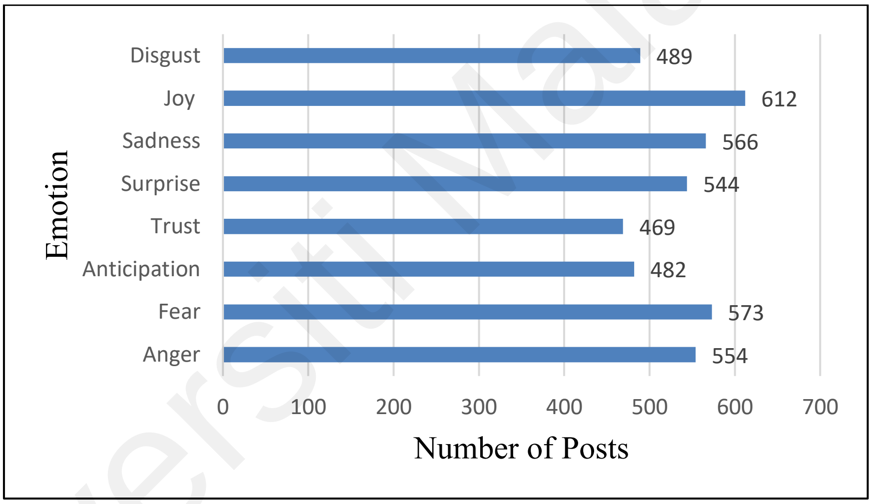


Figure 4.17 Number of posts per emotion

The contribution in this tier of classification is the inclusion of CBoW in order to determine the most dominant emotion unlike previous studies that have only worked on detecting the emotion within text (Chatterjee et al., 2019; Hasan et al., 2019; T. Rui, Cui, & Zhu, 2017).Therefore, results in this sub-section compares the proposed framework (M_{STEP-e}) against the benchmark, Emolex (M_{EMO}). Table 4.13 showcases the comparison results between the two aforementioned.

Table 4.12 Comparison Results between M_{EMO} and M_{STEP-e}

	Evaluation Metrics		
	F1-Score	AUC	Accuracy
M_{EMO}	0.69	0.63	0.64
M_{STEP-e}	**0.74**	**0.72**	**0.72**

M_{EMO} = Baseline without CBoW (common bag of words)
M_{STEP-e}*= Proposed framework with CBoW (common bag of words)*

Based on the results displayed above (Table 4.13), M_{STEP-e} outperformed M_{EMO} in terms of all the metrics used. Emolex detects emotions under the assumption that a post can have more than one emotion, however in classifying emotions, humans are able to identify the emotion which represents itself more strongly/dominantly after reading a post. Therefore, the effort to be able to replicate that ability was done within this research using the CBoW where the similarity check is carried out. The seed extension method used within this tier matches lexicon words within the training data with that from Emolex. This eventually enables the algorithm to determine the dominant emotion, and thus making it easier for emotion classifications to take place.

Table 4.13 Effectiveness based on Emotions

Emotion	Evaluation Metrics	
	F1 -Score	AUC
Anger	0.65	0.62
Fear	0.60	0.61
Anticipation	0.65	0.66
Trust	0.59	0.62
Surprise	0.60	0.61
Sadness	0.60	0.63
Joy	**0.68**	**0.67**
Disgust	0.54	0.59

Analysis on F1-Score and AUC for each emotion was also conducted with the results shown in Table 4.14. The lowest F1-Score was recorded at 0.54 for disgust. This could be due to the training size as seen in Figure 4.17, where the emotion disgust has the second lowest number of training data, thus affecting the classification effectiveness. Furthermore, there were a number of posts that were annotated as disgust; however, these posts were laced in sarcasm and irony. The proposed framework within this tier of classification does not cater for sarcasm and irony, hence emotion such as disgust that naturally lean towards sarcasm (Ravi & Ravi, 2017; Sulis, Irazú Hernández Farías, Rosso, Patti, & Ruffo, 2016) could not be classified as accurately.

STEP framework was able to classify anticipation and joy the best compared to other emotions. This gets contributed to the nature of the posts and how each user of the online health groups are all about spreading positivity (Maestre et al., 2018; McRoy et al., 2018; Rodrigues et al., 2016; Willis & Royne, 2017) hence the amount of posts that contain joy posts were larger. Generally, it can be concluded that the F1-Score remained within a range of 0.5 to 0.7 which can be considered an encouraging range despite the number of training data that was available for each emotion.

Other interesting findings that came across during the analysis phase were the use of verbs, adjectives and adverbs within the text that showed stronger emotions, and thus provoked users to click on the reaction buttons (love, wow, haha, sad, angry) on the posts more often. As discussed in sub-section 4.3.3, the love reaction was recorded the highest with 3,768 counts and haha was the lowest at 542 counts. However, the usage of the reaction haha was limited to the sense that it was only used in response to a post that showed some form of humor. The inference is that the nature of discussion that takes place within the group is more mature hence there is a lot of information exchange and less of conversational exchange.

Figure 4.18 shows the breakdown of emotions classified within the proposed STEP framework. When no emotion is detection, the particular post is categorized as No Emotion detected.

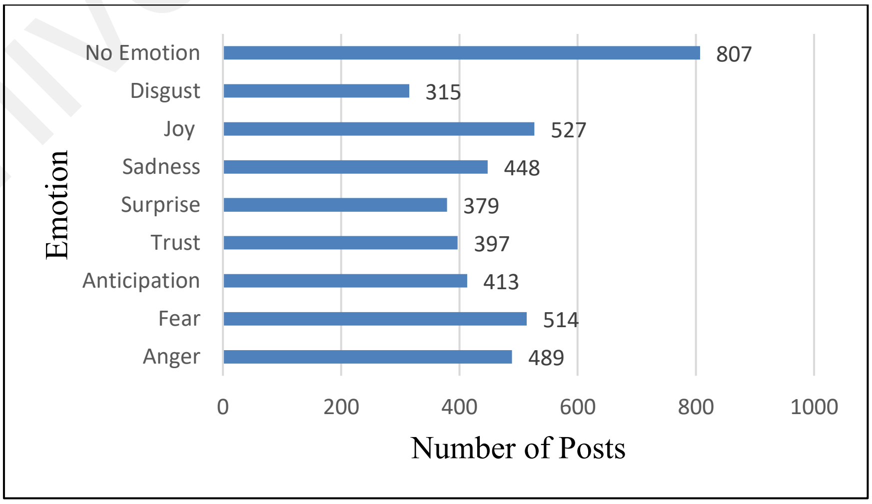


Figure 4.18 Emotion breakdown by posts after classification

In analyzing the emotions extracted from the posts, an experiment on the top most topics per emotion was conducted. Table 4.14 and Figure 4.19 depict the most discussed topics within the diabetes community extracted from the corpus.

Table 4.14 Top ten topics for each emotion

Sad		**Surprise**		**Joy**		**Fear**		**Anger**	
Words	***N***	***Words***	***N***	***Words***	***N***	***Words***	***N***	***Words***	***N***
Eat	117	OD	76	Can	136	Eat	130	Diabetes	57
Time	83	Eat	61	Eat	136	Can	121	Sugar	55
Sugar	80	Insulin	43	Diabetes	86	Diabetes	87	Help	45
Year	76	Diabetes	41	Insulin	87	Help	65	People	41
Help	69	Type	32	Carbs	77	People	53	Know	41
Carb	68	Carb	32	Good	77	Sugar	35	Years	39
People	49	Weight	31	Food	40	Blood	39	High	32
Blood	44	Diabetic	23	Sugars	32	Carbs	29	Blood	28
Water	17	Women	16	Water	13	Eating	29	Broken	16
Broken	16	Test	17	Meter	14	Food	25	Water	12

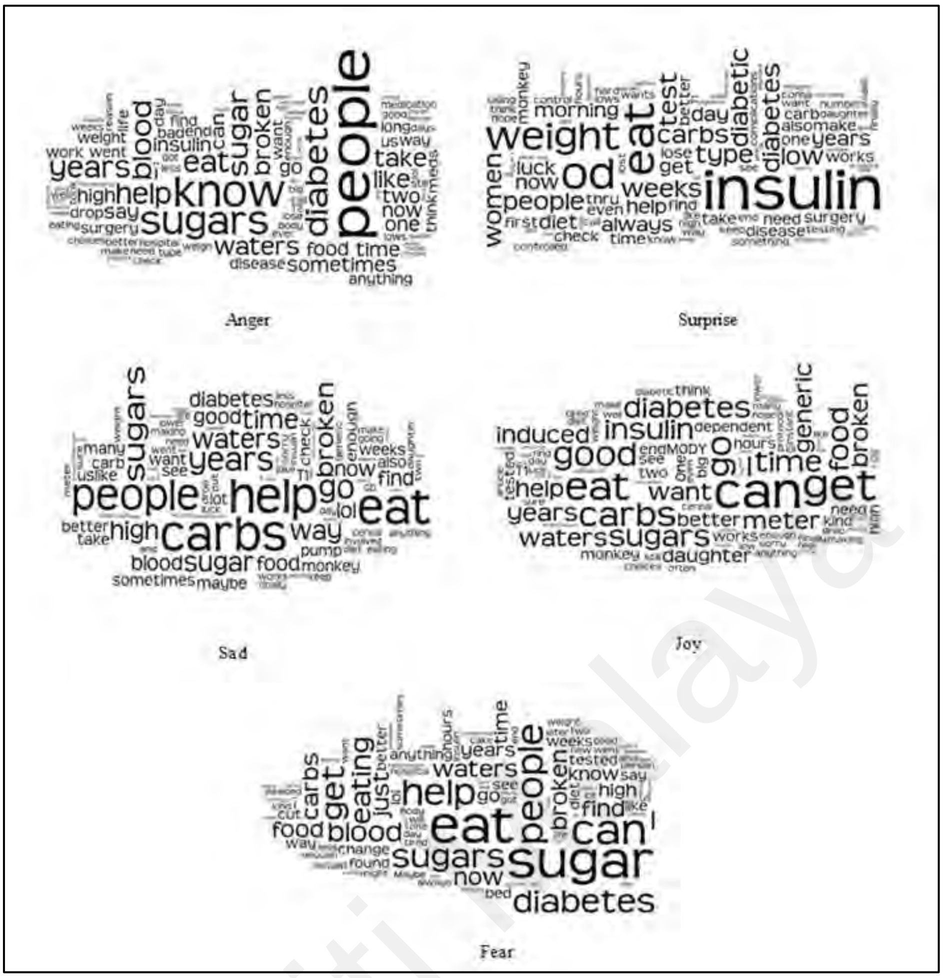


Figure 4.19 Top topics for top emotions

Based on the Figure 4.19, it shows words like *eat, diabetes, sugar, insulin* and *help* seem to be the common words used among the emotions displayed. For instance, sample posts related to the word *eat* can be seen as below:

Feeling guilty about your food choices could actually have a bigger impact on your metabolism than the food you're ***eat****ing!* - **Surprise**

I completely sympathize! I like eggs but not enough to ***eat*** *them every day! I also spike in the morning even with one piece of whole wheat toast* - **Sad**

Sugar free pudding, I used to ***eat*** *sugar chocolate pudding & mix some peanut butter with it and have a sliced apple to dip it in for a snack.* - **Joy**

Way toooo many carbs and sugars for someone with diabetes to ***eat*** *. . .* ***eat*** *this you'll have a major sugar spike . . . weigh the pros and cons is* ***eat****ing a piece of this cake worth losing a toe or your eyesight?* **- Fear**

Topics for joy were mostly related to a good diet, carbs, etc. suggesting that the majority of those emoting joy are probably sharing their successful recipes or lifestyle changes, and providing mental and emotional support to one another. This is very much in line with studies that found Facebook groups to serve as effective avenues to connect, share knowledge and provide peer support to each other (Oh et al., 2013; Y. Zhang et al., 2013). Sample posts below provide support to this finding:

How ***can*** *I help support a friend who is newly diagnosed with Type 2 diabetes?*

Only 6g of ***carbs*** *in this yummy dish! Enjoy it with some leftover Thanksgiving turkey instead of the sausage!*

When looking into the posts that have been identified as Joy, it was found that the use of emotionally linked words such as *happy*, *relieved* and *thankful* were mapped to the emotion Joy. And in many cases when one user posts a joyous post such as a successful delivery despite being diagnosed with gestational diabetes, many other users would in return convey their congratulations. On the other hand, the least number of posts were identified for Trust. This was an interesting discovery because most of the posts that actually showed the emotion Trust were from other users who were trying to comfort one another in making it through a difficult time while battling this disease. This is especially true for parents whose children were diagnosed with Type 1 diabetes.

4.4 Summary

This chapter presented the results produced by this research. Each result obtained was discussed individually using evaluation metrics as well as comparing it with benchmark models for verification and validation purposes.

For type classification, the contribution made was towards the manual lexicon dictionary that caters for all three types of diabetes. The type classification tier adopted the Naïve Bayes using n-gram and lexicon dictionary created. In comparing benchmark model against proposed STEP classification framework, the proposed framework performed better in classifying online health related data ($M_{Reichert}$) compared to M_{Salas} model. This goes to show the proposed lexicon dictionary was able to assist in improving the type classification process. Since STEP is a multi-tier framework, posts that were not able to be classified into a type were left as is and the already classified posts moved on to the next tier of purpose classification.

M_{STEP-p} used co-training Multinomial Naïve Bayes with weighted Information Gain feature selection to classify posts according to purpose labels of symptom, lifestyle and treatment. A dimension reduction technique of converting numeric vectors to string vectors was also performed in order to produce a better classification. This can be seen with respect to the results discussed. Apart from the standard evaluation metrics of F1-Score and AUC, this tier also tested the proposed classifier using Hamming Loss and 0/1 Loss and accuracy. Literature has stated these metrics are important to determine the efficiency of multi-label classification. Results show Hamming Loss results were lower than 0/1 Loss but this was due to the co-training algorithm adopted that treated the feature as input thus results of label classification outdid feature classification.

The next tier looked into sentiment and emotion classification. STEP framework includes sentiment intensity calculation with respect to the number of Facebook behaviors collected (like, comment, share and reaction). These nuances were converted to intensity using a proposed mathematical equation. Results show adding such intensity improved the classification process when comparing benchmark model against M_{STEP-S}. Emotion classification conducted also revealed interesting information that have been discussed in the sections above.

The overall STEP framework proposed is a multi-tier classification framework. With respect to the flow of data shown within the results discussion above, it is clear how redundant data is removed on the upper tiers leaving meaningful data for classification. Although the number of posts left for sentiment and emotion classification seems to be little, but those classifications were able to produce promising results as discussed. When only meaningful data gets classified into correct classification labels, the accuracy of the classifier is improved. The next chapter will look to provide a conclusion for this written thesis.

CHAPTER 5: CONCLUSION, LIMITATION AND FUTURE WORKS

This chapter serves as the conclusion chapter of this research. It will conclude the research as a whole and discuss limitations encountered along the way as well as identify areas within this research that can be looked into for future works.

The rest of this chapter is organized as follows: the overall research conclusion will be discussed in the first section, followed by the research contributions, research limitations and future works.

5.1 Research Conclusion

The existence of social media networks has created a platform for users across the globe to stay connected and share information with one another. The growing number of social media platforms such as Facebook, Twitter, Snapchat, Instagram etc. has only encouraged users to use such platforms to reach out to one another and share their life stories with their contacts (Groot et al., 2019; Poecze, Ebster, & Strauss, 2018; Roopchund et al., 2019). The aim of such platforms are not only to serve a single purpose or domain but has blossomed to include other areas such as medicine (Küçük, Yapar, Küçük, & Küçük, 2017; Martínez et al., 2016), businesses (Galati, Crescimanno, Tinervia, & Fagnani, 2017; Quesenberry & Coolsen, 2018), politics (Alashri et al., 2018; Sandoval-Almazan & Valle-Cruz, 2018) etc.

With the reservoir of textual data available, the study of mining opinion and analyzing sentiment could not have come at a better time. The study of sentiment has also expanded to include the study of emotions that can be extracted from text (Rout et al., 2018; Sailunaz et al., 2018). Therefore, the effort made in this research was to analyze text extracted from Facebook for not only sentiment and emotion but also to include purpose that proves to be useful in obtaining proper classification.

With a wide pool of data available, this research looked to limit its scope to posts extracted from Facebook from health-related groups. The decision to turn to diabetes came after consideration on the availability of the data and the significant impact this disease has not only on the growing population of Malaysia but also the Malaysian economy.

Through literature study it became apparent that many users were turning to online health support groups in order to seek support from others who were also battling the same disease or just to seek advice from others on the available treatment options (Abedin et al., 2017; Oh et al., 2013; Sharma et al., 2017). However, due to the vast availability of data, seeking the right information was an arduous task (Abedin et al., 2017). Therefore, the aim of this research was to automatically classify posts extracted from Facebook for sentiment, type, emotion and purpose in a manner that would improve classification process, thus providing users with information which caters best to their needs. Current studies that have looked to classify sentiment, emotion and purpose have done so in the field of politics (S. M. Mohammad et al., 2015) as well as within the domain of products and services (Al-Smadi et al., 2019; D. Anand & Naorem, 2016; Pham & Le, 2018; Poria, Cambria, & Gelbukh, 2016). However, the techniques used within these studies can be improved to produce better classifications. This led to identifying three objectives for this research:

1. To identify techniques and features to automatically classify posts based on Sentiment, Type, Emotion and Purpose.
2. To develop a multi-tier sentiment, type, emotion and purpose classification framework using sentiment intensity
3. To assess the proposed framework by means of experiments and evaluations.

With respect to the above, techniques that have been adopted for sentiment (Alashri et al., 2018; Sandoval-Almazan & Valle-Cruz, 2018; Verma & Thakur, 2018), emotion (Chatzakou, Vakali, & Kafetsios, 2017; Rout et al., 2017; Waterloo, Baumgartner, Peter, & Valkenburg, 2017; Yadollahi et al., 2017) and purpose (S. M. Mohammad et al., 2015) classifications in the past were studied extensively. These were discussed in detail in Chapter 2 where several gaps within the studies were identified. One of the problem statements identified was the classification accuracy achieved in the past with respect to each separate element (sentiment, emotion, purpose). Therefore, this research adopted different techniques within different tiers that would help boost the classification process, and thus improving the classification accuracy. These included using weighted information gain feature selection method that reassigns weights on wrongly classified data, hence improving the training process through the cross validation fold, mathematically calculating sentiment intensity with respect to the number of features extracted from Facebook (like, comment, share and reaction) and also using common bag of words method to detect the dominant emotion within a text in order to improve emotion classification.

Furthermore, techniques that were identified in past literature either used machine learning, lexicon based or a combination of the two. As the research progressed, it became apparent that this research had to implement different techniques within each tier as the classification within each tier varied from one another. Features that could also improve the classification of sentiment, emotion and purpose were also examined. This research discovered Facebook features such as number of likes, comment, shares and reaction can actually contribute towards sentiment detection (C. Kim & Yang, 2017; Quesenberry & Coolsen, 2018), however studies thus far has yet to manipulate these features as a measure of opinion strength. Therefore, the present research collected the number of likes, comments, shares and reactions were obtained for each post, and later assigned different weights to each of these features with respect to its significance within literature. These were presented and elaborated in Chapter 4.

The second objective was to develop the proposed multi-tier sentiment, type, emotion and purpose (STEP) classification framework. To recap, a manually created lexicon dictionary along with Naïve Bayes algorithm were used in the first tier (Type classification). This was then followed by the purpose classification whereby a string-based Multinomial Naïve Bayes was adapted to include co-training as well as weighted information gain feature selection technique to achieve a better classification result. The third tier classified sentiments by including Facebook features (like, comment, share, reaction) which were converted into sentiment intensity counts before finally performing emotion classification.

The final objective was to assess the proposed framework, which was done separately due to the different techniques applied within each tier. To be specific, this was accomplished using evaluation metrics (F1-Score, area under curve and accuracy) and benchmark dataset comparisons, both of which described in Chapter 3. The results of the evaluation phase were presented and discussed in Chapter 4. In this chapter, it was found that the proposed framework produced more accurate classifications compared to benchmark studies, thus validating the techniques adopted. Specifically, the proposed framework produced 77% F1-Score compared to the benchmark (i.e. 70%) for type classification. Similarly, the inclusion of co-training along with a weighted feature selection technique and conversion of string vectors, improved the F1-Score from 38% to 61%. The inclusion of Facebook behaviors was also able to improve the accuracy from 64% to 77%, similar with emotion detection where a baseline F1-Score of 69% improved to 74% with the use of common bag of words.

In conclusion, the objectives of this research have been met, and each has been discussed in length within the chapters of this thesis. The results that were discussed in Chapter 4 show the techniques and features identified were able to improve the classification method, and thus showing the proposed framework is able to classify sentiment, emotion and purpose more accurately.

5.2 Research Contributions

The contributions of this research are as follows:

- A multi-tier sentiment, type, emotion, purpose (STEP) classification framework

  The proposed framework of including all elements within a single framework has not been carried out till date. Each tier of STEP framework caters for a different need, and the classification in each tier contributes to refine the training data by removing redundant posts that do not contribute towards classification. Similar to aspect classification of products and services, the type and purpose classification tier of STEP framework groups aspects of text extracted from social media into groups for classification. This allows for better structure to emerge from unstructured social media text which then allows users to look out for information in a much easier manner compared to scavenging through thousands of posts.

- The use of Facebook features (like, comment, share and reaction) to improve sentiment classification.

  This research has also contributed a sentiment intensity calculation which takes Facebook features into consideration. Previous studies have only looked into the content analysis perspective (Househ, 2016; McRoy et al., 2018; Y. Zhang et al., 2013); however, this research used those studies as a foundation and proposed to include those nuances as intensity, and thus improving the overall sentiment classification.

- Using weighted information gain feature selection and string-based vectors to reduce dimensionality and improve classification accuracy.

  The algorithm proposed in the purpose classification tier adopted weighted feature selection method that learns from wrongly classified text and re-assigns its weight to further improve classification. This is done in addition to co-training the algorithm which adds feature classification into label classification as input, resulting in better results. Furthermore, in an effort to reduce dimensionality and sparse distribution of features within the classification process, numerical vectors were converted to string vectors.

- A manual lexicon dictionary for type classification catering to all three types of diabetes was also created.

  A lexicon dictionary for type of diabetes classification was created. Literature study has revealed one diabetes ontology which caters for type 1 diabetes (El-Sappagh & Ali, 2016). A lexicon dictionary that also caters for type 2 and gestational diabetes did not exist. Furthermore, the aforementioned ontology was more inclined towards medical diagnosis instead of general symptoms and treatment. Therefore, another contribution of this research was the creation of a lexicon dictionary.

5.3 Research Significance

Pharmaceutical companies would also benefit from this classification framework as they would be able to use it to assess the feedback on medication they are looking to import or even medication they have just distributed amongst health practitioners within the country. The continuous improvement and release of new drugs in combatting diseases means pharmaceutical companies need to always be in the know of new drugs or drugs that have caused negative reactions from the public. The purpose classification of this framework for example, enables the companies to focus on the treatment classification and to easily gauge users' perceptions (i.e. sentiment and emotion).

Additionally, the proposed classification framework is scalable to cater for other diseases, such as cancer, mental illness, genetic diseases etc. The potential of this framework is not only limited within the medical domain as it can be also extended to support business sectors, such as movie, hotel and even airline reviews.

5.4 Research Limitations and Future Works

As with majority of research, the presented work of this thesis comes with its own limitations. The first limitation lies in the diabetes dataset itself. The dataset comprised of posts that were only in English, and thus the framework would not be able to perform well for posts in other languages. Future studies could look into expanding the framework to support other popular languages such as Arabic and Spanish as well as other datasets apart from diabetes.

The next limitation comes in the form of catering to irony and sarcasm. Sarcasm and irony are both defined as a negative sentiment disguised as a positive sentiment (Mukherjee & Bala, 2017; Ravi & Ravi, 2017). This research took each post at face value without checking for sarcasm or irony which could lead to a different sentiment score as well as emotion. This is because, similar to sentiment, a sarcastic post could come across as a *joy* but in actual manner it could be of *anger*. For future works, the inclusion of sarcasm and irony can be looked into in order to improve the sentiment and emotion classifications.

Finally, another limitation lies within the processing time of the framework itself. Although the existence of multiple tiers allows redundant and irrelevant data to be removed in the upper tiers as proved in literature (Baqapuri et al., 2016; Xu et al., 2015) as well as this research, it also poses another limitation on the processing time it takes for the framework to produce classification results. On average, the time taken from data pre-processing to emotion classification is approximately 0.8 hours (48 minutes). This may not be feasible for a classification to take place in a real time setup. One of the possibilities of future works is to adopt other forms of classification such as AdaBoost or ensemble classifiers that would be able to improve the classification time.

REFERENCES

Abedin, T., Al Mamun, M., Lasker, M. A. A., Ahmed, S. W., Shommu, N., Rumana, N., & Turin, T. C. (2017). Social Media as a Platform for Information About Diabetes Foot Care: A Study of Facebook Groups. *Canadian Journal of Diabetes, 41*(1), 97-101.

Agarwal, S., & Sureka, A. (2016). *But i did not mean it!—intent classification of racist posts on tumblr.* Paper presented at the Intelligence and Security Informatics Conference (EISIC), 2016 European.

Ahmad, F. K. (2017). Comparative Analysis of Feature Extraction Techniques for Event Detection from News Channels' Facebook Page. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 9*(1-2), 13-17.

Aisopos, F., Tzannetos, D., Violos, J., & Varvarigou, T. (2016, March 29 2016-April 1 2016). *Using N-Gram Graphs for Sentiment Analysis: An Extended Study on Twitter.* Paper presented at the 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService).

Akaichi, J. (2014). *Sentiment classification at the time of the tunisian uprising: Machine learning techniques applied to a new corpus for arabic language.* Paper presented at the Proceedings - 2014 European Network Intelligence Conference, ENIC 2014.

Akaichi, J., Dhouioui, Z., & Lopez-Huertas Perez, M. J. (2013). *Text mining facebook status updates for sentiment classification.* Paper presented at the 2013 17th International Conference on System Theory, Control and Computing, ICSTCC 2013; Joint Conference of SINTES 2013, SACCS 2013, SIMSIS 2013 - Proceedings.

Akhtar, M. S., Gupta, D., Ekbal, A., & Bhattacharyya, P. (2017). Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis. *Knowledge-Based Systems, 125*, 116-135.

Akter, S., & Aziz, M. T. (2016). *Sentiment analysis on facebook group using lexicon based approach.* Paper presented at the Electrical Engineering and Information Communication Technology (ICEEICT), 2016 3rd International Conference on.

Al-Anzi, F. S., & AbuZeina, D. (2017). Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. *Journal of King Saud University-Computer and Information Sciences, 29*(2), 189-195.

Al-Smadi, M., Al-Ayyoub, M., Jararweh, Y., & Qawasmeh, O. (2019). Enhancing aspect-based sentiment analysis of Arabic hotels' reviews using morphological, syntactic and semantic features. *Information Processing & Management, 56*(2), 308-319.

Alahmari, S., & Buckley, J. (2015). *Sentiment analysis in Arabic tweets.* Paper presented at the IMSCI 2015 - 9th International Multi-Conference on Society, Cybernetics and Informatics, Proceedings.

Alashri, S., Srivatsav Kandala, S., Bajaj, V., Parriott, E., Awazu, Y., & C Desouza, K. (2018). *The 2016 US Presidential Election on Facebook: An Exploratory Analysis of Sentiments.* Paper presented at the Proceedings of the 51st Hawaii International Conference on System Sciences.

Aldoğan, D., & Yaslan, Y. (2017). A comparison study on active learning integrated ensemble approaches in sentiment analysis. *Computers & Electrical Engineering, 57*, 311-323.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.

AlQarni, Z. A., Yunus, F., & Househ, M. S. (2016). Health information sharing on Facebook: an exploratory study on diabetes mellitus. *Journal of infection and public health, 9*(6), 708-712.

Alsaedi, N., Burnap, P., & Rana, O. (2017). Can we predict a riot? disruptive event detection using twitter. *ACM Transactions on Internet Technology (TOIT), 17*(2), 18.

Altınel, B., Can Ganiz, M., & Diri, B. (2017). Instance labeling in semi-supervised learning with meaning values of words. *Engineering Applications of Artificial Intelligence, 62*, 152-163.

Altınel, B., & Ganiz, M. C. (2016). A new hybrid semi-supervised algorithm for text classification with class-based semantics. *Knowledge-Based Systems, 108*, 50-64.

Altrabsheh, N., Cocea, M., & Fallahkhair, S. (2014). Learning sentiment from students' feedback for real-time interventions in classrooms. In *Adaptive and Intelligent Systems* (pp. 40-49): Springer.

An, Y., Sun, S., & Wang, S. (2017). *Naive Bayes classifiers for music emotion classification based on lyrics.* Paper presented at the 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS).

Anand, D., & Naorem, D. (2016). Semi-supervised aspect based sentiment analysis for movies using review filtering. *Procedia Computer Science, 84*, 86-93.

Anand, N., Goyal, D., & Kumar, T. (2018). *Analyzing and Preprocessing the Twitter Data for Opinion Mining.* Paper presented at the Proceedings of International Conference on Recent Advancement on Computer and Communication.

Andreu-Perez, J., Poon, C. C., Merrifield, R. D., Wong, S. T., & Yang, G.-Z. (2015). Big data for health. *IEEE journal of biomedical and health informatics, 19*(4), 1193-1208.

Appel, O., Chiclana, F., Carter, J., & Fujita, H. (2016). A hybrid approach to the sentiment analysis problem at the sentence level. *Knowledge-Based Systems, 108*, 110-124.

Aravindan, S., & Ekbal, A. (2014). *Feature extraction and opinion mining in online product reviews.* Paper presented at the Proceedings - 2014 13th International Conference on Information Technology, ICIT 2014.

Association, A. D. (2005). Diagnosis and classification of diabetes mellitus. *Diabetes Care, 28*(1), S37.

Association, A. D. (2010). Diagnosis and classification of diabetes mellitus. *Diabetes Care, 33*(Supplement 1), S62-S69.

Association, A. D. (2018). Economic Costs of Diabetes in the US in 2017. *Diabetes Care, 41*(5), 917-928.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). *Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.* Paper presented at the LREC.

Bagheri, A., & Saraee, M. (2014). Persian sentiment analyzer: A framework based on a novel feature selection method. *International Journal of Artificial Intelligence, 12*(2), 115-129.

Bagheri, A., Saraee, M., & de Jong, F. (2013). Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. *Knowledge-Based Systems, 52*, 201-213.

Balahur, A., & Turchi, M. (2012). *Comparative experiments for multilingual sentiment analysis using machine translation.* Paper presented at the CEUR Workshop Proceedings.

Balakrishnan, V., & Lloyd-Yemoh, E. (2014). Stemming and lemmatization: a comparison of retrieval performances. *Lecture Notes on Software Engineering, 2*(3), 262.

Balazs, J. A., & Velásquez, J. D. (2016). Opinion Mining and Information Fusion: A survey. *Information Fusion, 27*, 95-110.

Baqapuri, A. I., Saleh, S., Ilyas, M. U., Khan, M. M., & Qamar, A. M. (2016, 22-27 May 2016). *Sentiment classification of tweets using hierarchical classification.* Paper presented at the 2016 IEEE International Conference on Communications (ICC).

Barnaghi, P., Ghaffari, P., & Breslin, J. G. (2016, March 29 2016-April 1 2016). *Opinion Mining and Sentiment Polarity on Twitter and Correlation between Events and Sentiment.* Paper presented at the 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService).

Bazarova, N. N., Choi, Y. H., Schwanda Sosik, V., Cosley, D., & Whitlock, J. (2015). *Social sharing of emotions on Facebook: Channel differences, satisfaction, and replies.* Paper presented at the Proceedings of the 18th ACM conference on computer supported cooperative work & social computing.

Bilal, M., Israr, H., Shahid, M., & Khan, A. (2016). Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques. *Journal of King Saud University - Computer and Information Sciences, 28*(3), 330-344.

Bilici, E., & Saygın, Y. (2017). Why do people (not) like me?: Mining opinion influencing factors from reviews. *Expert Systems with Applications, 68*, 185-195.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145-1159.

Bravo-Marquez, F., Frank, E., & Pfahringer, B. (2016). Building a Twitter opinion lexicon from automatically-annotated tweets. *Knowledge-Based Systems, 108*, 65-78.

Breen, J. O. (2012). Mining twitter for airline consumer sentiment. *Practical text mining and statistical analysis for non-structured text data applications, 133*.

Bu, Z., Li, H., Cao, J., Wu, Z., & Zhang, L. (2016). Game theory based emotional evolution analysis for chinese online reviews. *Knowledge-Based Systems, 103*, 60-72.

Burnap, P., Rana, O. F., Avis, N., Williams, M., Housley, W., Edwards, A., . . . Sloan, L. (2015). Detecting tension in online communities with computational Twitter analysis. *Technological Forecasting and Social Change, 95*, 96-108.

Calero, A. (2013). Likes vs. comments vs. shares. *Retrieved April, 26*, 2015.

Calheiros, A. C., Moro, S., & Rita, P. (2017). Sentiment classification of consumer-generated online reviews using topic modeling. *Journal of Hospitality Marketing & Management, 26*(7), 675-693.

Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems, 31*(2), 102-107.

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). *A practical guide to sentiment analysis* (Vol. 5): Springer.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems, 28*(2), 15-21.

Canales, L., Strapparava, C., Boldrini, E., & Martnez-Barco, P. (2016). *Exploiting a bootstrapping approach for automatic annotation of emotions in texts.* Paper presented at the Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on.

Carah, N. (2014). Like, comment, share: Alcohol brand activity on Facebook.

Chakraborty, K., Bhattacharyya, S., Bag, R., & Hassanien, A. E. (2018). *Comparative sentiment analysis on a set of movie reviews using deep learning approach.* Paper presented at the International Conference on Advanced Machine Learning Technologies and Applications.

Chandra Pandey, A., Singh Rajpoot, D., & Saraswat, M. (2017). Twitter sentiment analysis using hybrid cuckoo search method. *Information Processing & Management, 53*(4), 764-779.

Chang, C.-C., Chiu, S.-I., & Hsu, K.-W. (2017). *Predicting political affiliation of posts on Facebook.* Paper presented at the Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication.

Charalampakis, B., Spathis, D., Kouslis, E., & Kermanidis, K. (2016). A comparison between semi-supervised and supervised text mining techniques on detecting irony in greek political tweets. *Engineering Applications of Artificial Intelligence, 51*, 50-57.

Chatrath, A., Miao, H., Ramchander, S., & Villupuram, S. (2014). Currency jumps, cojumps and the role of macro news. *Journal of International Money and Finance, 40*, 42-62.

Chatterjee, A., Narahari, K. N., Joshi, M., & Agrawal, P. (2019). *SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text.* Paper presented at the Proceedings of the 13th International Workshop on Semantic Evaluation.

Chatzakou, D., Vakali, A., & Kafetsios, K. (2017). Detecting variation of emotions in online activities. *Expert Systems with Applications, 89*, 318-332.

Chaudhuri, A. (2006). *Emotion and reason in consumer behavior*: Routledge.

Chen, J., Chen, H., Wu, Z., Hu, D., & Pan, J. Z. (2017). Forecasting smog-related health hazard based on social media and physical sensor. *Information Systems, 64*, 281-291.

Chen, T., Xu, R., He, Y., & Wang, X. (2017). Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications, 72*, 221-230.

Chifu, E. Ş., Leţia, T. Ş., & Chifu, V. R. (2015). *Unsupervised aspect level sentiment analysis using Ant Clustering and Self-organizing Maps.* Paper presented at the Speech Technology and Human-Computer Dialogue (SpeD), 2015 International Conference on.

Claypo, N., & Jaiyen, S. (2014). *Opinion mining for Thai restaurant reviews using neural networks and mRMR feature selection.* Paper presented at the 2014 International Computer Science and Engineering Conference, ICSEC 2014.

Corazza, M., Menini, S., Arslan, P., Sprugnoli, R., Cabrio, E., Tonelli, S., & Villata, S. (2018). *Comparing Different Supervised Approaches to Hate Speech Detection.* Paper presented at the EVALITA 2018.

Cosma, G., & Acampora, G. (2016). A computational intelligence approach to efficiently predicting review ratings in e-commerce. *Applied Soft Computing, 44*, 153-162.

Coursaris, C. K., van Osch, W., & Balogh, B. A. (2016). *Do facebook likes lead to shares or sales? exploring the empirical links between social media content, brand* *equity, purchase intention, and engagement.* Paper presented at the System Sciences (HICSS), 2016 49th Hawaii International Conference on.

Cvijikj, I. P., & Michahelles, F. (2013). Online engagement factors on Facebook brand pages. *Social Network Analysis and Mining, 3*(4), 843-861.

da Silva, N. F. F., Coletta, L. F. S., Hruschka, E. R., & Hruschka Jr, E. R. (2016). Using unsupervised information to improve semi-supervised tweet sentiment classification. *Information Sciences, 355–356*, 348-365.

Dai, H., & Hao, J. (2017). Mining social media data on marijuana use for Post Traumatic Stress Disorder. *Computers in Human Behavior, 70*, 282-290.

Das, B., & Prathima, V. R. (2016). A Framework to Formulate Customer Taste from Unstructured Review Data. *Procedia Computer Science, 93*, 1046-1053.

Deng, Z.-H., Luo, K.-H., & Yu, H.-L. (2014). A study of supervised term weighting scheme for sentiment analysis. *Expert Systems with Applications, 41*(7), 3506-3513.

Desmet, B., & Hoste, V. (2013). Emotion detection in suicide notes. *Expert Systems with Applications, 40*(16), 6351-6358.

Devika, M. D., Sunitha, C., & Ganesh, A. (2016). Sentiment Analysis: A Comparative Study on Different Approaches. *Procedia Computer Science, 87*, 44-49.

Dhaoui, C., Webster, C. M., & Tan, L. P. (2017). Social media sentiment analysis: lexicon versus machine learning. *Journal of Consumer Marketing, 34*(6), 480-488.

Dhar, A., Mukherjee, H., Dash, N. S., & Roy, K. (2018). *Performance of Classifiers in Bangla Text Categorization.* Paper presented at the 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET).

Diab, D. M., & El Hindi, K. M. (2017). Using differential evolution for fine tuning naïve Bayesian classifiers and its application for text classification. *Applied Soft Computing, 54*, 183-199.

Ding, C., Cheng, H. K., Duan, Y., & Jin, Y. (2017). The power of the “like” button: The impact of social media on box office. *Decision Support Systems, 94*, 77-84.

Divya, R., Sandhya, S., & Sai, V. S. (2015). Sentiment Analysis and Effective Visualization of Faculty and Course Feedback. *The International Journal of Science and Technoledge, 3*(5), 125.

Du, Y., Liu, J., Ke, W., & Gong, X. (2018). Hierarchy construction and text classification based on the relaxation strategy and least information model. *Expert systems with applications, 100*, 157-164.

Duwairi, R. M., & Qarqaz, I. (2014). *Arabic sentiment analysis using supervised classification.* Paper presented at the Proceedings - 2014 International Conference on Future Internet of Things and Cloud, FiCloud 2014.

Ekström, M., & Östman, J. (2015). Information, interaction, and creative production: The effects of three forms of internet use on youth democratic engagement. *Communication Research, 42*(6), 796-818.

El-Sappagh, S., & Ali, F. (2016). *DDO: a diabetes mellitus diagnosis ontology.* Paper presented at the Applied Informatics.

Elghazel, H., Aussem, A., Gharroudi, O., & Saadaoui, W. (2016). Ensemble multi-label text categorization based on rotation forest and latent semantic indexing. *Expert Systems with Applications, 57*, 1-11.

Elhadad, M. K., Li, K. F., & Gebali, F. (2019). *Sentiment Analysis of Arabic and English Tweets.* Paper presented at the Workshops of the International Conference on Advanced Information Networking and Applications.

Fergie, G., Hilton, S., & Hunt, K. (2016). Young adults' experiences of seeking online information about diabetes and mental health in the age of social media. *Health Expectations, 19*(6), 1324-1335.

Fernández-Gavilanes, M., Álvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E., & Javier González-Castaño, F. (2016). Unsupervised method for sentiment analysis in online texts. *Expert Systems with Applications, 58*, 57-75.

Fersini, E., Messina, E., & Pozzi, F. A. (2016). Expressive signals in social media languages to improve polarity detection. *Information Processing & Management, 52*(1), 20-35.

Frank, E., & Bouckaert, R. R. (2006). *Naive bayes for text classification with unbalanced classes.* Paper presented at the European Conference on Principles of Data Mining and Knowledge Discovery.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning, 29*(2-3), 131-163.

Frison, E., & Eggermont, S. (2015). The impact of daily stress on adolescents' depressed mood: The role of social support seeking through Facebook. *Computers in Human Behavior, 44*, 315-325.

Fu, P.-W., Wu, C.-C., & Cho, Y.-J. (2017). What makes users share content on facebook? Compatibility among psychological incentive, social capital focus, and content type. *Computers in Human Behavior, 67*, 23-32.

Fu, X., Li, J., Yang, K., Cui, L., & Yang, L. (2016). Dynamic Online HDP model for discovering evolutionary topics from Chinese social texts. *Neurocomputing, 171*, 412-424.

Gage-Bouchard, E. A., LaValley, S., Mollica, M., & Beaupin, L. K. (2017). Cancer communication on social media: Examining how cancer caregivers use Facebook for cancer-related communication. *Cancer nursing, 40*(4), 332-338.

Galati, A., Crescimanno, M., Tinervia, S., & Fagnani, F. (2017). Social media as a strategic marketing tool in the Sicilian wine industry: Evidence from Facebook. *Wine Economics and Policy, 6*(1), 40-47.

Ganeshbhai, S. Y., & Shah, B. K. (2015). *Feature based opinion mining: A survey.* Paper presented at the Advance Computing Conference (IACC), 2015 IEEE International.

Gao, K., Xu, H., & Wang, J. (2015). *Emotion cause detection for chinese micro-blogs based on ECOCC model.* Paper presented at the Pacific-Asia Conference on Knowledge Discovery and Data Mining.

García-Pablos, A., Cuadros, M., & Rigau, G. (2018). W2VLDA: almost unsupervised system for aspect based sentiment analysis. *Expert Systems with Applications, 91*, 127-137.

Genc-Nayebi, N., & Abran, A. (2017). A systematic literature review: Opinion mining studies from mobile app store user reviews. *Journal of Systems and Software, 125*, 207-219.

Gezici, G., Yanikoglu, B., Tapucu, D., & Saygin, Y. (2012). *New features for sentiment analysis: Do sentences matter?* Paper presented at the CEUR Workshop Proceedings.

Ghazi, D., Inkpen, D., & Szpakowicz, S. (2010). *Hierarchical versus flat classification of emotions in text.* Paper presented at the Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text.

Ghorpade, T., & Ragha, L. (2012). *Featured based sentiment classification for hotel reviews using NLP and Bayesian classification.* Paper presented at the Proceedings - 2012 International Conference on Communication, Information and Computing Technology, ICCICT 2012.

Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G., & Chatzisavvas, K. C. (2017). Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications, 69*, 214-224.

Godbole, S., & Sarawagi, S. (2004). *Discriminative methods for multi-labeled classification.* Paper presented at the Pacific-Asia conference on knowledge discovery and data mining.

Greene, J. A., Choudhry, N. K., Kilabuk, E., & Shrank, W. H. (2011). Online social networking by patients with diabetes: a qualitative evaluation of communication with Facebook. *Journal of general internal medicine, 26*(3), 287-292.

Groot, P., Westermann-Behaylo, M. K., Rehbein, K., & Perrault Crawford, E. (2019). *Stakeholder Activism and Corporate Environmental Performance: The Role of Social Media.* Paper presented at the Academy of Management Proceedings.

Group, N. D. D. (1979). Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance. *Diabetes, 28*(12), 1039-1057.

Gulaty, M. (2016). *Aspect-Based Sentiment Analysis.* Dublin, National College of Ireland,

Gull, R., Shoaib, U., Rasheed, S., Abid, W., & Zahoor, B. (2016). Pre Processing of Twitter's Data for Opinion Mining in Political Context. *Procedia Computer Science, 96*, 1560-1570.

Guo, L., Shi, F., & Tu, J. (2016). Textual analysis and machine leaning: Crack unstructured data in finance and accounting. *The Journal of Finance and Data Science, 2*(3), 153-170.

Habernal, I., Ptáček, T., & Steinberger, J. (2014). Supervised sentiment analysis in Czech social media. *Information Processing and Management, 50*(5), 693-707. doi:10.1016/j.ipm.2014.05.001

Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science, 17*, 26-32.

Hammami, S. (2016). The Three Phases of Facebook: Social Networks and the Public Sphere in the Arab World—the Case of the Tunisian Revolution. In *Media and Political Contestation in the Contemporary Arab World* (pp. 35-61): Springer.

Harpel, T. (2018). Pregnant women sharing pregnancy-related information on Facebook: web-based survey study. *Journal of Medical Internet Research, 20*(3), e115.

Harrag, F., & Al-Qawasmah, E. (2010). Improving Arabic Text Categorization Using Neural Network with SVD. *JDIM, 8*(4), 233-239.

Hasan, M., Rundensteiner, E., & Agu, E. (2019). Automatic emotion detection in text streams by analyzing Twitter data. *International Journal of Data Science and Analytics, 7*(1), 35-51.

Hashimoto, K., Kontonatsios, G., Miwa, M., & Ananiadou, S. (2016). Topic detection using paragraph vectors to support active learning in systematic reviews. *Journal of Biomedical Informatics, 62*, 59-65.

Hays, S., Page, S. J., & Buhalis, D. (2013). Social media as a destination marketing tool: its use by national tourism organisations. *Current issues in Tourism, 16*(3), 211-239.

Heston, S. L., & Sinha, N. R. (2014). News versus sentiment: Comparing textual processing approaches for predicting stock returns. *Robert H. Smith School Research Paper*.

Hodeghatta, U. R. (2013). *Sentiment analysis of Hollywood movies on Twitter.* Paper presented at the Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2013.

Hong, S., & Cameron, G. T. (2018). Will comments change your opinion? The persuasion effects of online comments and heuristic cues in crisis communication. *Journal of Contingencies and Crisis Management, 26*(1), 173-182.

Househ, M. (2016). Communicating Ebola through social media and electronic news media outlets: A cross-sectional study. *Health informatics journal, 22*(3), 470-478.

Hu, M., & Liu, B. (2004). *Mining and summarizing customer reviews.* Paper presented at the Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.

Ibrahim, O. A., & Landa-Silva, D. (2014). *A new weighting scheme and discriminative approach for information retrieval in static and dynamic document collections.* Paper presented at the 2014 14th UK Workshop on Computational Intelligence (UKCI).

Idrees, F., Rajarajan, M., Conti, M., Chen, T. M., & Rahulamathavan, Y. (2017). PIndroid: A novel Android malware detection system using ensemble learning methods. *Computers & Security, 68*, 36-46.

Isguder-Sahin, G. G., Zafer, H. R., & Adah, E. (2014). *Polarity detection of Turkish comments on technology companies.* Paper presented at the Proceedings of the International Conference on Asian Language Processing 2014, IALP 2014.

Islam, M. S., Islam, M. A., Hossain, M. A., & Dey, J. J. (2016). *Supervised approach of sentimentality extraction from bengali facebook status.* Paper presented at the Computer and Information Technology (ICCIT), 2016 19th International Conference on.

Ive, J., Gkotsis, G., Dutta, R., Stewart, R., & Velupillai, S. (2018). *Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health.* Paper presented at the Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic.

Jabbar, A., Iqbal, S., Khan, M. U. G., & Hussain, S. (2018). A survey on Urdu and Urdu like language stemmers and stemming techniques. *Artificial Intelligence Review, 49*(3), 339-373.

Jagtap, M. S. N., & Adamuthe, A. C. (2018). Hierarchical Text Classification using Dictionary-Based Approach and Convolutional Neural Network.

Jain, A., & Mandowara, J. (2016). Text classification by combining text classifiers to improve the efficiency of classification. *International Journal of Computer Application (2250-1797), 6*(2).

Jeyapriya, A., & Selvi, C. S. K. (2015). *Extracting aspects and mining opinions in product reviews using supervised learning algorithm.* Paper presented at the 2nd International Conference on Electronics and Communication Systems, ICECS 2015.

Ji, X., Chun, S. A., & Geller, J. (2013). *Monitoring public health concerns using twitter sentiment classifications.* Paper presented at the Proceedings - 2013 IEEE International Conference on Healthcare Informatics, ICHI 2013.

Jiang, D., Tao, Q., Wang, Z., & Dong, L. (2019). An Intelligent Logistic Regression Approach for Verb Expression's Sentiment Analysis. In *Recent Developments in Intelligent Computing, Communication and Devices* (pp. 173-181): Springer.

Jiang, H., Qiang, M., & Lin, P. (2016). Assessment of online public opinions on large infrastructure projects: A case study of the Three Gorges Project in China. *Environmental Impact Assessment Review, 61*, 38-51.

Jiang, L., Li, C., Wang, S., & Zhang, L. (2016). Deep feature weighting for naive Bayes and its application to text classification. *Engineering Applications of Artificial Intelligence, 52*, 26-39.

Jiang, L., Wang, S., Li, C., & Zhang, L. (2016). Structure extended multinomial naive Bayes. *Information Sciences, 329*, 346-356.

Jiang, L., Zhang, L., Li, C., & Wu, J. (2018). A Correlation-based Feature Weighting Filter for Naive Bayes. *IEEE Transactions on Knowledge and Data Engineering*.

Jijkoun, V., de Rijke, M., & Weerkamp, W. (2010). *Generating focused topic-specific sentiment lexicons.* Paper presented at the Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.

Jiménez-Zafra, S. M., Martín-Valdivia, M. T., Molina-González, M. D., & Ureña-López, L. A. (2019). How do we talk about doctors and drugs? Sentiment analysis in forums expressing opinions for medical domain. *Artificial Intelligence in Medicine, 93*, 50-57.

Jindal, N., & Liu, B. (2006). *Mining comparative sentences and relations.* Paper presented at the AAAI.

Jo, T. (2016). *Encoding Words into String Vectors for Word Categorization.* Paper presented at the Proceedings on the International Conference on Artificial Intelligence (ICAI).

Jo, T. (2017a). *Long text segmentation by string vector based KNN.* Paper presented at the Advanced Communication Technology (ICACT), 2017 19th International Conference on.

Jo, T. (2017b). *String vector based AHC for text clustering.* Paper presented at the Advanced Communication Technology (ICACT), 2017 19th International Conference on.

Jo, T., & Cho, D.-H. (2008). Index based approach for text categorization. *International journal of mathematics and computers in simulation, 1*(2), 127-132.

Jotheeswaran, J., & Koteeswaran, S. (2016). Feature selection using random forestmethod for sentiment analysis. *Indian Journal of Science and Technology, 9*(3), 1-7. doi:10.17485/ijst/2016/v9i3/86387

Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review, 52*(1), 273-292.

Kaur, H. J., & Kumar, R. (2015). *Sentiment analysis from social media in crisis situations.* Paper presented at the International Conference on Computing, Communication and Automation, ICCCA 2015.

Kaur, W., & Balakrishnan, V. (2018). Improving sentiment scoring mechanism: a case study on airline services. *Industrial Management & Data Systems, 118*(8), 1578-1596.

Kaur, W., Balakrishnan, V., Rana, O., & Sinniah, A. (2019). Liking, sharing, commenting and reacting on Facebook: User behaviors' impact on sentiment intensity. *Telematics and Informatics, 39*, 25-36.

Khan, A. U. R., Khan, M., & Khan, M. B. (2016). Naïve Multi-label Classification of YouTube Comments Using Comparative Opinion Mining. *Procedia Computer Science, 82*, 57-64.

Khan, F. H., Qamar, U., & Bashir, S. (2016a). eSAP: A decision support framework for enhanced sentiment analysis and polarity classification. *Information Sciences, 367–368*, 862-873

Khan, F. H., Qamar, U., & Bashir, S. (2016b). SWIMS: Semi-supervised subjective feature weighting and intelligent model selection for sentiment analysis. *Knowledge-Based Systems, 100*, 97-111.

Kim, C., & Yang, S.-U. (2017). Like, comment, and share on Facebook: How each behavior differs from the other. *Public Relations Review, 43*(2), 441-449.

Kim, E. H.-J., Jeong, Y. K., Kim, Y., Kang, K. Y., & Song, M. (2016). Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news. *Journal of Information Science, 42*(6), 763-781.

Kim, H., Howland, P., & Park, H. (2005). Dimension reduction in text classification with support vector machines. *Journal of Machine Learning Research, 6*(Jan), 37-53.

Kim, S., Zhang, J., Chen, Z., Oh, A. H., & Liu, S. (2013). *A Hierarchical Aspect-Sentiment Model for Online Reviews.* Paper presented at the AAAI.

Korkontzelos, I., Nikfarjam, A., Shardlow, M., Sarker, A., Ananiadou, S., & Gonzalez, G. H. (2016). Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *Journal of Biomedical Informatics, 62*, 148-158.

Kowsari, K., Brown, D. E., Heidarysafa, M., Meimandi, K. J., Gerber, M. S., & Barnes, L. E. (2017). *Hdltex: Hierarchical deep learning for text classification.* Paper presented at the Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on.

Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J., & Woźniak, M. (2017). Ensemble learning for data stream analysis: A survey. *Information Fusion, 37*, 132-156.

Krebs, F., Lubascher, B., Moers, T., Schaap, P., & Spanakis, G. (2017). Social Emotion Mining Techniques for Facebook Posts Reaction Prediction. *arXiv preprint arXiv:1712.03249*.

Krippendorff, K. (2004). Reliability in content analysis. *Human communication research, 30*(3), 411-433.

Krippendorff, K. (2011). Computing Krippendorff's alpha-reliability.

Krouska, A., Troussas, C., & Virvou, M. (2016). *The effect of preprocessing techniques on Twitter sentiment analysis.* Paper presented at the 2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA).

Küçük, E. E., Yapar, K., Küçük, D., & Küçük, D. (2017). Ontology-based automatic identification of public health-related Turkish tweets. *Computers in Biology and Medicine, 83*, 1-9.

Kumar, A., Kansal, C., & Ekbal, A. (2015). *Investigating active learning techniques for document level sentiment classification of tweets.* Paper presented at the 2015 7th International Conference on Communication Systems and Networks, COMSNETS 2015 - Proceedings.

Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems, 114*, 128-147.

Le, H. S., Van Le, T., & Pham, T. V. (2015). *Aspect Analysis for Opinion Mining of Vietnamese Text.* Paper presented at the Advanced Computing and Applications (ACOMP), 2015 International Conference on.

Lee, C.-H. (2018). Multi-label classification of documents using fine-grained weights and modified co-training. *Intelligent Data Analysis, 22*(1), 103-115.

Lee, J., & Kim, D.-W. (2017). SCLS: Multi-label feature selection based on scalable criterion for large label set. *Pattern Recognition, 66*, 342-352.

Lek, H. H., & Poo, D. C. C. (2013). *Aspect-based Twitter sentiment classification.* Paper presented at the Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI.

Li, G. (2017). *Application of sentiment analysis: assessing the reliability and validity of the global airlines rating program.* University of Twente,

Li, J., Fong, S., Zhuang, Y., & Khoury, R. (2016). Hierarchical classification in text mining for sentiment analysis of online news. *Soft Computing, 20*(9), 3411-3420.

Li, J., Rao, Y., Jin, F., Chen, H., & Xiang, X. (2016). Multi-label maximum entropy model for social emotion classification over short text. *Neurocomputing, 210*, 247-256.

Li, W., & Xu, H. (2014). Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications, 41*(4), 1742-1749.

Liao, C., Feng, C., Yang, S., & Huang, H. (2016). Topic-related Chinese message sentiment analysis. *Neurocomputing, 210*, 237-246.

Liau, B. Y., & Tan, P. P. (2014). Gaining customer knowledge in low cost airlines through text mining. *Industrial Management and Data Systems, 114*(9), 1344-1359. doi:10.1108/IMDS-07-2014-0225

Lim, K. W., Buntine, W., Chen, C., & Du, L. (2016). Nonparametric Bayesian topic modelling with the hierarchical Pitman–Yor processes. *International Journal of Approximate Reasoning, 78*, 172-191.

Lim, S., Tucker, C. S., & Kumara, S. (2017). An unsupervised machine learning model for discovering latent infectious diseases using social media data. *Journal of Biomedical Informatics, 66*, 82-94.

Lima, A. C. E. S., De Castro, L. N., & Corchado, J. M. (2015). A polarity analysis framework for Twitter messages. *Applied Mathematics and Computation, 270*, 756-767. doi:10.1016/j.amc.2015.08.059

Lin, H.-C. (2017). How political candidates' use of Facebook relates to the election outcomes. *International Journal of Market Research, 59*(1), 77-96.

Lin, Y.-S., Jiang, J.-Y., & Lee, S.-J. (2014). A similarity measure for text classification and clustering. *IEEE Transactions on Knowledge and Data Engineering, 26*(7), 1575-1590.

Liu, B. (2007). Web usage mining. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, 449-483.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies, 5*(1), 1-167.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*: Cambridge University Press.

Liu, S. M., & Chen, J.-H. (2015). A multi-label classification based approach for sentiment classification. *Expert Systems with Applications, 42*(3), 1083-1093.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing, 234*, 11-26.

Liu, Y., Bi, J.-W., & Fan, Z.-P. (2017). Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms. *Expert Systems with Applications, 80*, 323-339.

Liu, Z., Liu, S., Liu, L., Sun, J., Peng, X., & Wang, T. (2016). Sentiment recognition of online course reviews using multi-swarm optimization-based selected features. *Neurocomputing, 185*, 11-20.

Lo, S. L., Cambria, E., Chiong, R., & Cornforth, D. (2016). A multilingual semi-supervised approach in deriving Singlish sentic patterns for polarity detection. *Knowledge-Based Systems, 105*, 236-247.

Lochter, J. V., Zanetti, R. F., Reller, D., & Almeida, T. A. (2016). Short text opinion detection using ensemble of classifiers and semantic indexing. *Expert Systems with Applications, 62*, 243-249.

Lu, Y., Wu, Y., Liu, J., Li, J., & Zhang, P. (2017). Understanding health care social media use from different stakeholder perspectives: a content analysis of an online health community. *Journal of Medical Internet Research, 19*(4).

Ma, J., & Hinrichs, E. (2015). *Accurate linear-time Chinese word segmentation via embedding matching.* Paper presented at the Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). *Learning word vectors for sentiment analysis.* Paper presented at the Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1.

Maestre, J., Herring, S., Min, A., Connelly, C., & Shih, P. (2018). Where and How to Look for Help Matters: Analysis of Support Exchange in Online Health Communities for People Living with HIV. *Information, 9*(10), 259.

Malik, H., & Shakshuki, E. M. (2016). Mining Collective Opinions for Comparison of Mobile Apps. *Procedia Computer Science, 94*, 168-175.

Malik, M. S. I., & Hussain, A. (2017). Helpfulness of product reviews as a function of discrete positive and negative emotions. *Computers in Human Behavior, 73*, 290-302.

Martínez, P., Martínez, J. L., Segura-Bedmar, I., Moreno-Schneider, J., Luna, A., & Revert, R. (2016). Turning user generated health-related content into actionable knowledge through text analytics services. *Computers in Industry, 78*, 43-56.

Mazzonello, V., Gaglio, S., Augello, A., & Pilato, G. (2013). *A study on classification methods applied to sentiment analysis.* Paper presented at the Proceedings - 2013 IEEE 7th International Conference on Semantic Computing, ICSC 2013.

McRoy, S., Rastegar-Mojarad, M., Wang, Y., Ruddy, K. J., Haddad, T. C., & Liu, H. (2018). Assessing unmet information needs of breast cancer survivors: Exploratory study of online health forums using text classification and retrieval. *JMIR cancer, 4*(1), e10.

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

Meire, M., Ballings, M., & Van den Poel, D. (2016). The added value of auxiliary data in sentiment analysis of Facebook posts. *Decision Support Systems, 89*, 98-112.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39-41.

Mirończuk, M. M., & Protasiewicz, J. (2018). A Recent Overview of the State-of-the-Art Elements of Text Classification. *Expert Systems with Applications*.

Mishra, D., Venugopalan, M., & Gupta, D. (2016). Context Specific Lexicon for Hindi Reviews. *Procedia Computer Science, 93*, 554-563.

Moh, M., Gajjala, A., Gangireddy, S. C. R., & Moh, T. S. (2015, 6-9 Dec. 2015). *On Multi-tier Sentiment Analysis Using Supervised Machine Learning.* Paper presented at the 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT).

Mohammad, S., & Bravo-Marquez, F. (2017). *Emotion Intensities in Tweets.* Paper presented at the Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017).

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence, 29*(3), 436-465.

Mohammad, S. M., Zhu, X., Kiritchenko, S., & Martin, J. (2015). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management, 51*(4), 480-499.

Moraes, R., Valiati, J. F., & Neto, W. P. G. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications, 40*(2), 621-633.

Mountassir, A., Benbrahim, H., & Berrada, I. (2012). *Addressing the problem of unbalanced data sets in sentiment analysis.* Paper presented at the KDIR 2012 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval.

Mujtaba, G., Shuib, L., Raj, R. G., Al-Garadi, M. A., Rajandram, R., & Shaikh, K. (2017). *Hierarchical Text Classification of Autopsy Reports to Determine MoD and CoD through Term-Based and Concepts-Based Features.* Paper presented at the Industrial Conference on Data Mining.

Mukherjee, S., & Bala, P. K. (2017). Sarcasm detection in microblogs using Naïve Bayes and fuzzy clustering. *Technology in Society, 48*, 19-27.

Mummalaneni, V., Gruss, R., Goldberg, D. M., Ehsani, J. P., & Abrahams, A. S. (2018). Social media analytics for quality surveillance and safety hazard detection in baby cribs. *Safety Science, 104*, 260-268.

Nagamma, P., Pruthvi, H. R., Nisha, K. K., & Shwetha, N. H. (2015). *An improved sentiment analysis of online movie reviews based on clustering for box-office prediction.* Paper presented at the International Conference on Computing, Communication and Automation, ICCCA 2015.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2015). Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. *Expert Systems with Applications, 42*(1), 306-324.

Natarajan, P., Sankaran, V., Santhi, B., & Brindha, G. R. (2013). Polarity categorization with fine tuned pipeline process of online reviews. *International Journal of Engineering and Technology, 5*(3), 2221-2226.

Nayak, A., Pai, M. M. M., & Pai, R. M. (2016). Prediction Models for Indian Stock Market. *Procedia Computer Science, 89*, 441-449.

Nguyen, D. Q., Vu, T., Nguyen, D. Q., Dras, M., & Johnson, M. (2017). From Word Segmentation to POS Tagging for Vietnamese. *arXiv preprint arXiv:1711.04951*.

Nguyen, T. D., Tran, T., Phung, D., & Venkatesh, S. (2016). Graph-induced restricted Boltzmann machines for document modeling. *Information Sciences, 328*, 60-75

Oeldorf-Hirsch, A., & Sundar, S. S. (2015). Posting, commenting, and tagging: Effects of sharing news stories on Facebook. *Computers in Human Behavior, 44*, 240-249.

Oh, H. J., Lauckner, C., Boehmer, J., Fewins-Bliss, R., & Li, K. (2013). Facebooking for health: An examination into the solicitation and effects of health-related social support on social networking sites. *Computers in Human Behavior, 29*(5), 2072-2080.

Okada, M., Takeuchi, K., & Hashimoto, K. (2014). *A method to classify customer reviews of Japanese hotels by support vector machine using estimation sentence patterns information.* Paper presented at the Proceedings - 2014 IIAI 3rd International Conference on Advanced Applied Informatics, IIAI-AAI 2014.

Onan, A., Korukoğlu, S., & Bulut, H. (2017). A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. *Information Processing & Management, 53*(4), 814-833.

Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior, 31*, 527-541.

Pak, Y., Kim, Y. J., Song, M., & Kim, Y. H. (2015). Shock waves of political risk on the stock market: The case of Korean companies in the U.S. *Development and Society, 44*(1), 143-165.

Paltoglou, G., & Thelwall, M. (2017). Sensing social media: a range of approaches for sentiment analysis. In *Cyberemotions* (pp. 97-117): Springer.

Parackal, M., Parackal, S., Eusebius, S., & Mather, D. (2017). The Use of Facebook Advertising for Communicating Public Health Messages: A Campaign Against Drinking During Pregnancy in New Zealand. *JMIR public health and surveillance, 3*(3), e49.

Parkhe, V., & Biswas, B. (2014). *Aspect based sentiment analysis of movie reviews: Finding the polarity directing aspects.* Paper presented at the Proceedings - 2014 International Conference on Soft Computing and Machine Intelligence, ISCMI 2014.

Patra, B. G., Mukherjee, N., Das, A., Mandal, S., Das, D., & Bandyopadhyay, S. (2014). *Identifying aspects and analyzing their sentiments from reviews.* Paper presented at the Proceedings of Special Session 2014 13th Mexican International Conference on Artificial Intelligence: Advances in Artificial Intelligence, MICAI 2014.

Pelletier, M. J., & Blakeney Horky, A. (2015). Exploring the Facebook Like: a product and service perspective. *Journal of Research in Interactive Marketing, 9*(4), 337-354.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates, 71*(2001), 2001.

Perikos, I., & Hatzilygeroudis, I. (2016). Recognizing emotions in text using ensemble of classifiers. *Engineering Applications of Artificial Intelligence, 51*, 191-201.

Pham, D.-H., & Le, A.-C. (2018). Learning multiple layers of knowledge representation for aspect based sentiment analysis. *Data & Knowledge Engineering, 114*, 26-39.

Pirrelli, V., & Zarghili, A. (2017). Arabic Natural Language Processing: Models, systems and applications. *Journal of King Saud University - Computer and Information Sciences, 29*(2), A1-A3.

Plutchik, R. (2003). *Emotions and life: Perspectives from psychology, biology, and evolution*: American Psychological Association.

Poecze, F., Ebster, C., & Strauss, C. (2018). Social media metrics and sentiment analysis to evaluate the effectiveness of social media posts. *Procedia Computer Science, 130*, 660-666.

Pool, C., & Nissim, M. (2016). Distant supervision for emotion detection using Facebook reactions. *arXiv preprint arXiv:1611.02988*.

Poria, S., Cambria, E., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems, 108*, 42-49.

Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., & Morency, L.-P. (2017). *Context-dependent sentiment analysis in user-generated videos.* Paper presented at the Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Poria, S., Cambria, E., Howard, N., Huang, G.-B., & Hussain, A. (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing, 174*, 50-59.

Pröllochs, N., Feuerriegel, S., & Neumann, D. (2015). *Enhancing sentiment analysis of financial news by detecting negation scopes.* Paper presented at the System Sciences (HICSS), 2015 48th Hawaii International Conference on.

Purohit, H., Dong, G., Shalin, V., Thirunarayan, K., & Sheth, A. (2015). *Intent classification of short-text on social media.* Paper presented at the Smart

City/SocialCom/SustainCom (SmartCity), 2015 IEEE International Conference on.

Qiu, G., Liu, B., Bu, J., & Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational Linguistics, 37*(1), 9-27.

Quan, C., & Ren, F. (2010). A blog emotion corpus for emotional expression analysis in Chinese. *Computer Speech & Language, 24*(4), 726-749.

Quesenberry, K. A., & Coolsen, M. K. (2018). What Makes Facebook Brand Posts Engaging? A Content Analysis of Facebook Brand Post Text That Increases Shares, Likes, and Comments to Influence Organic Viral Reach. *Journal of Current Issues & Research in Advertising*, 1-16.

Rahmawati, D., & Khodra, M. L. (2016). *Word2vec semantic representation in multilabel classification for Indonesian news article.* Paper presented at the 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA).

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems, 89*, 14-46.

Ravi, K., & Ravi, V. (2017). A novel automatic satire and irony detection using ensembled feature selection and data mining. *Knowledge-Based Systems, 120*, 15-33.

Razzaq, M. A., Qamar, A. M., & Bilal, H. S. M. (2014). *Prediction and analysis of Pakistan election 2013 based on sentiment analysis.* Paper presented at the ASONAM 2014 - Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning, 85*(3), 333.

Rehman, A., Javed, K., & Babri, H. A. (2017). Feature selection based on a normalized difference measure for text classification. *Information Processing & Management, 53*(2), 473-489.

Reichert, J.-R., Kristensen, K. L., Mukkamala, R. R., & Vatrapu, R. (2017). *A supervised machine learning study of online discussion forums about type-2 diabetes.* Paper presented at the 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom).

Ristoski, P., & Paulheim, H. (2016). Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Web Semantics: Science, Services and Agents on the World Wide Web, 36*, 1-22.

Rodrigues, R. G., das Dores, R. M., Camilo-Junior, C. G., & Rosa, T. C. (2016). SentiHealth-Cancer: A sentiment analysis tool to help detecting mood of patients in online social networks. *International Journal of Medical Informatics, 85*(1), 80-95.

Rohani, V. A., & Shayaa, S. (2015). *Utilizing machine learning in Sentiment Analysis: SentiRobo approach.* Paper presented at the 2nd International Symposium on Technology Management and Emerging Technologies, ISTMET 2015 - Proceeding.

Roopchund, R., Ramesh, V., & Jaunky, V. (2019). Use of Social Media for Improving Student Engagement at Université des Mascareignes (UDM). In *Information Systems Design and Intelligent Applications* (pp. 11-20): Springer.

Ross, B., Potthoff, T., Majchrzak, T. A., Chakraborty, N. R., Ben Lazreg, M., & Stieglitz, S. (2018). *The Diffusion of Crisis-Related Communication on Social Media: An Empirical Analysis of Facebook Reactions.* Paper presented at the Proceedings of the 51st Hawaii International Conference on System Sciences.

Rosso, P., Bosco, C., Damiano, R., Patti, V., & Cambria, E. (2016). Emotion and sentiment in social and expressive media: Introduction to the special issue. *Information Processing & Management, 52*(1), 1-4.

Rout, J. K., Choo, K.-K. R., Dash, A. K., Bakshi, S., Jena, S. K., & Williams, K. L. (2017). A model for sentiment and emotion analysis of unstructured social media text. *Electronic Commerce Research*, 1-19.

Rout, J. K., Choo, K.-K. R., Dash, A. K., Bakshi, S., Jena, S. K., & Williams, K. L. (2018). A model for sentiment and emotion analysis of unstructured social media text. *Electronic Commerce Research, 18*(1), 181-199.

Rui, J. R., & Stefanone, M. A. (2013). Strategic image management online: Self-presentation, self-esteem and social network perspectives. *Information, Communication & Society, 16*(8), 1286-1305.

Rui, T., Cui, P., & Zhu, W. (2017). Joint user-interest and social-influence emotion prediction for individuals. *Neurocomputing, 230*, 66-76.

Rus, H. M., & Cameron, L. D. (2016). Health communication in social media: message features predicting user engagement on diabetes-related Facebook pages. *Annals of Behavioral Medicine, 50*(5), 678-689.

Ruuska, S., Hämäläinen, W., Kajava, S., Mughal, M., Matilainen, P., & Mononen, J. (2018). Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle. *Behavioural processes, 148*, 56-62.

Sailunaz, K., Dhaliwal, M., Rokne, J., & Alhajj, R. (2018). Emotion detection from text and speech: a survey. *Social Network Analysis and Mining, 8*(1), 28.

Salas-Zárate, M. d. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodriguez-Garcia, M. A., & Valencia-García, R. (2017). Sentiment analysis on tweets about diabetes: an aspect-level approach. *Computational and mathematical methods in medicine, 2017*.

Salas-Zárate, M. d. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodríguez-García, M. Á., & Valencia-García, R. (2017). Sentiment analysis on tweets about diabetes: an aspect-level approach. *Computational and mathematical methods in medicine, 2017*.

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2017). A survey of text mining in social media: facebook and twitter perspectives. *Adv. Sci. Technol. Eng. Syst. J, 2*(1), 127-133.

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613-620.

Samal, B., Behera, A. K., & Panda, M. (2017). *Performance analysis of supervised machine learning techniques for sentiment analysis.* Paper presented at the Sensing, Signal Processing and Security (ICSSS), 2017 Third International Conference on.

Sandoval-Almazan, R., & Valle-Cruz, D. (2018). *Facebook impact and sentiment analysis on political campaigns.* Paper presented at the Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age.

Seerat, B., & Azam, F. (2012). Opinion mining: Issues and challenges (a survey). *International Journal of Computer Applications, 49*(9).

Setyawan, M. Y. H., Awangga, R. M., & Efendi, S. R. (2018). *Comparison Of Multinomial Naive Bayes Algorithm And Logistic Regression For Intent Classification In Chatbot.* Paper presented at the 2018 International Conference on Applied Engineering (ICAE).

Shamim, A. (2015). *Improving online decision making process based on the ranking of user reviews and product features.* University of Malaya,

Sharma, M., Yadav, K., Yadav, N., & Ferdinand, K. C. (2017). Zika virus pandemic—analysis of Facebook as a social media health information platform. *American journal of infection control, 45*(3), 301-302.

Sherawat, D. (2015). *Mining emotions (anger & fear) from Indian army fans page on Facebook.* Paper presented at the Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances in.

Shoukry, A., & Rafea, A. (2012). *Sentence-level Arabic sentiment analysis.* Paper presented at the Proceedings of the 2012 International Conference on Collaboration Technologies and Systems, CTS 2012.

Silla, C. N., & Freitas, A. A. (2011). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery, 22*(1-2), 31-72.

Singh, T., & Kumari, M. (2016). Role of Text Pre-processing in Twitter Sentiment Analysis. *Procedia Computer Science, 89*, 549-554.

Smieško, I. (2016). Criminal Liability of Facebook Reaction Buttons in Environment of the Slovak Republic as a Form of Hate Speech. *SOCIETAS ET IURISPRUDENTIA, 4*, 213.

Sokolova, M., & Bobicev, V. (2013). *What sentiments can be found in medical forums?* Paper presented at the International Conference Recent Advances in Natural Language Processing, RANLP.

Song, P., Zheng, W., Ou, S., Zhang, X., Jin, Y., Liu, J., & Yu, Y. (2016). Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization. *Speech Communication, 83*, 34-41.

Srisuan, J., & Hanskunatai, A. (2014). *The ensemble of Naïve Bayes classifiers for hotel searching.* Paper presented at the 2014 International Computer Science and Engineering Conference, ICSEC 2014.

Stone, P., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1968). The general inquirer: A computer approach to content analysis. *Journal of Regional Science, 8*(1), 113-116.

Sulis, E., Irazú Hernández Farías, D., Rosso, P., Patti, V., & Ruffo, G. (2016). Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems, 108*, 132-143.

Sumner, E. M., Ruge-Jones, L., & Alcorn, D. (2017). A functional approach to the Facebook Like button: An exploration of meaning, interpersonal functionality, and potential alternative response buttons. *New Media & Society*, 1461444817697917.

Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion, 36*, 10-25.

Sun, X., Zhang, C., Li, G., Sun, D., Ren, F., Zomaya, A., & Ranjan, R. (2018). Detecting users' anomalous emotion using social media for business intelligence. *Journal of Computational Science, 25*, 193-200.

Sundermann, C. V., Domingues, M. A., Conrado, M. d. S., & Rezende, S. O. (2016). Privileged contextual information for context-aware recommender systems. *Expert Systems with Applications, 57*, 139-158.

Swain, S., & Seeja, K. (2017). *Analysis of Epidemic Outbreak in Delhi Using Social Media Data.* Paper presented at the International Conference on Information, Communication and Computing Technology.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*(2), 267-307.

Teh, P. L., Pak, I., Rayson, P., & Piao, S. (2015). *Exploring fine-grained sentiment values in online product reviews.* Paper presented at the Open Systems (ICOS), 2015 IEEE Confernece on.

Tellez, E. S., Miranda-Jiménez, S., Graff, M., Moctezuma, D., Siordia, O. S., & Villaseñor, E. A. (2017). A case study of Spanish text transformations for twitter sentiment analysis. *Expert Systems with Applications, 81*, 457-471.

Thelwall, M. (2013). Heart and soul: Sentiment strength detection in the social web with sentistrength. *Proceedings of the CyberEmotions*, 1-14.

Tian, Y., Galery, T., Dulcinati, G., Molimpakis, E., & Sun, C. (2017). Facebook Sentiment: Reactions and Emojis. *SocialNLP 2017*, 11.

Tian, Y., Zhang, Y., & Liu, D. (2016). Semi-supervised support vector classification with self-constructed Universum. *Neurocomputing, 189*, 33-42.

Toujani, R., & Akaichi, J. (2017). *Fuzzy sentiment classification in social network facebook'statuses mining.* Paper presented at the Information and Digital Technologies (IDT), 2017 International Conference on.

Trinh, S., Nguyen, L., Vo, M., & Do, P. (2016). Lexicon-based sentiment analysis of Facebook comments in Vietnamese language. In *Recent Developments in Intelligent Information and Database Systems* (pp. 263-276): Springer.

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

Troussas, C., Virvou, M., Espinosa, K. J., Llaguno, K., & Caro, J. (2013, 10-12 July 2013). *Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning.* Paper presented at the 2013 Fourth International Conference on Information, Intelligence, Systems and Applications (IISA).

Tsur, G., Pinter, Y., Szpektor, I., & Carmel, D. (2016). *Identifying web queries with question intent.* Paper presented at the Proceedings of the 25th International Conference on World Wide Web.

Turnbull, S., & Jenkins, S. (2016). Why Facebook Reactions are good news for evaluating social media campaigns. *Journal of Direct, Data and Digital Marketing Practice, 17*(3), 156-158.

Uysal, A. K. (2016). An improved global feature selection scheme for text classification. *Expert Systems with Applications, 43*, 82-92.

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management, 50*(1), 104-112.

Varghese, R., & Jayasree, M. (2013). *Aspect based Sentiment Analysis using support vector machine classifier.* Paper presented at the Proceedings of the 2013 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2013.

Vepsäläinen, T., Li, H., & Suomi, R. (2017). Facebook likes and public opinion: Predicting the 2015 Finnish parliamentary elections. *Government Information Quarterly*.

Verma, B., & Thakur, R. S. (2018). *Sentiment Analysis Using Lexicon and Machine Learning-Based Approaches: A Survey.* Paper presented at the Proceedings of International Conference on Recent Advancement on Computer and Communication.

Vidya, N. A., Fanany, M. I., & Budi, I. (2015). Twitter Sentiment to Analyze Net Brand Reputation of Mobile Phone Providers. *Procedia Computer Science, 72*, 519-526.

Vilares, D., Alonso, M. A., & Gómez-Rodríguez, C. (2017). Supervised sentiment analysis in multilingual environments. *Information Processing & Management, 53*(3), 595-607.

Vilares, J., Vilares, M., Alonso, M. A., & Oakes, M. P. (2016). On the feasibility of character n-grams pseudo-translation for Cross-Language Information Retrieval tasks. *Computer Speech & Language, 36*, 136-164.

Wang, G., Zhang, Z., Sun, J., Yang, S., & Larson, C. A. (2015). POS-RS: A Random Subspace method for sentiment classification based on part-of-speech analysis. *Information Processing and Management, 51*(4), 458-479. doi:10.1016/j.ipm.2014.09.004

Wang, W., & Wang, H. (2015). Opinion-enhanced collaborative filtering for recommender systems through sentiment analysis. *New Review of Hypermedia and Multimedia, 21*(3-4), 278-300.

Wang, Y., Huang, J., He, T., & Tu, X. (2019). Dialogue intent classification with character-CNN-BGRU networks. *Multimedia Tools and Applications*, 1-20.

Waterloo, S. F., Baumgartner, S. E., Peter, J., & Valkenburg, P. M. (2017). Norms of online expressions of emotion: Comparing Facebook, Twitter, Instagram, and WhatsApp. *New Media & Society*, 1461444817707349.

Williams, T. P., & Gong, J. (2014). Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. *Automation in Construction, 43*, 23-29.

Willis, E., & Royne, M. B. (2017). Online health communities and chronic disease self-management. *Health communication, 32*(3), 269-278.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). *Recognizing contextual polarity in phrase-level sentiment analysis.* Paper presented at the Proceedings of the conference on human language technology and empirical methods in natural language processing.

Wongso, R., Luwinda, F. A., Trisnajaya, B. C., & Rusli, O. (2017). News article text classification in Indonesian language. *Procedia Computer Science, 116*, 137-143.

Wu, L., Moh, T. S., & Khuri, N. (2015, Oct. 29 2015-Nov. 1 2015). *Twitter opinion mining for adverse drug reactions.* Paper presented at the Big Data (Big Data), 2015 IEEE International Conference on.

Xu, H., Yang, W., & Wang, J. (2015). Hierarchical emotion classification and emotion component analysis on Chinese micro-blog posts. *Expert systems with applications, 42*(22), 8745-8752.

Xue, B., Fu, C., & Shaobin, Z. (2014). *A study on sentiment computing and classification of sina weibo with word2vec.* Paper presented at the 2014 IEEE International Congress on Big Data.

Yadollahi, A., Shahraki, A. G., & Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR), 50*(2), 25.

Yan, Z., Wang, T., Chen, Y., & Zhang, H. (2016). Knowledge sharing in online health communities: A social exchange theory perspective. *Information & Management, 53*(5), 643-653.

Yang, B., & Cardie, C. (2014). *Context-aware learning for sentence-level sentiment analysis with posterior regularization.* Paper presented at the Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Yang, X., Li, G., & Huang, S. S. (2017). Perceived online community support, member relations, and commitment: differences between posters and lurkers. *Information & Management, 54*(2), 154-165.

Yoon, H. G., Kim, H., Kim, C. O., & Song, M. (2016). Opinion polarity detection in Twitter data combining shrinkage regression and topic modeling. *Journal of Informetrics, 10*(2), 634-644.

Zell, A. L., & Moeller, L. (2018). Are you happy for me… on Facebook? The potential importance of “likes” and comments. *Computers in Human Behavior, 78*, 26-33.

Zhai, C. (2017). *Probabilistic topic models for text data retrieval and analysis.* Paper presented at the Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval.

Zhang, D., Xu, H., Su, Z., & Xu, Y. (2015). Chinese comments sentiment classification based on word2vec and SVMperf. *Expert Systems with Applications, 42*(4), 1857-1863.

Zhang, H. (2004). The optimality of naive Bayes. *AA, 1*(2), 3.

Zhang, H., Sekhari, A., Ouzrout, Y., & Bouras, A. (2016). Jointly identifying opinion mining elements and fuzzy measurement of opinion intensity to analyze product features. *Engineering Applications of Artificial Intelligence, 47*, 122-139.

Zhang, L., & Liu, B. (2017). Sentiment analysis and opinion mining. In *Encyclopedia of Machine Learning and Data Mining* (pp. 1152-1161): Springer.

Zhang, Y., He, D., & Sang, Y. (2013). Facebook as a platform for health information and communication: a case study of a diabetes group. *Journal of medical systems, 37*(3), 9942.

Zhao, J., Liu, K., & Xu, L. (2016). Sentiment analysis: mining opinions, sentiments, and emotions. In: MIT Press.

Zhao, K., Yen, J., Greer, G., Qiu, B., Mitra, P., & Portier, K. (2014). Finding influential users of online health communities: a new metric based on sentiment influence. *Journal of the American Medical Informatics Association, 21*(e2), e212-e218.

Zhao, Y., Dong, S., & Yang, J. (2015). Effect research of aspects extraction for Chinese hotel reviews based on machine learning method. *International Journal of Smart Home, 9*(3), 23-34. doi:10.14257/ijsh.2015.9.3.03

Zhou, Y., Tong, Y., Gu, R., & Gall, H. (2014). *Combining Text Mining and Data Mining for Bug Report Classification.* Paper presented at the 2014 IEEE International Conference on Software Maintenance and Evolution.