IMPROVING MULTI-LABEL TEXT CLASSIFICATION USING WEIGHTED INFORMATION GAIN AND CO-TRAINED MULTINOMIAL NAÏVE BAYES CLASSIFIER

Wandeep Kaur[1], Vimala Balakrishnan[2*] and Kok-Seng Wong [3]

[1]Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, 43600, Bangi, Malaysia

[2]Department of Information Systems, Faculty of Computer Science and Information Technology, Universiti Malaya, 50603, Kuala Lumpur, Malaysia.

[3]College of Engineering and Computer Science, VinUniversity, Vinhomes Ocean Park, 100000 Hanoi, Vietnam.

Email: wandeep@ukm.edu.my[1], vimala.balakrishnan@um.edu.my[2*] (corresponding author), wong.ks@vinuni.edu.vn [3]

DOI: https://doi.org/10.22452/mjcs.vol35no1.2

ABSTRACT

Over recent years, the emergence of electronic text processing systems has generated a vast amount of structured and unstructured data, thus creating a challenging situation for users to rummage through irrelevant information. Therefore, studies are continually looking to improve the classification process to produce more accurate results that would benefit users. This paper looks into the weighted information gain method that re-assigns wrongly classified features with new weights to provide better classification. The method focuses on the weights of the frequency bins, assuming every time a certain word frequency bin is iterated, it provides information on the target word feature. Therefore, the more iteration and re-assigning of weight occur within the bin, the more important the bin becomes, eventually providing better classification. The proposed algorithm was trained and tested using a corpus extracted from dedicated Facebook pages related to diabetes. The weighted information gain feature selection technique is then fed into a co-trained Multinomial Naïve Bayes classification algorithm that captures the labels' dependencies. The algorithm incorporates class value dependencies since the dataset used multi-label data before converting string vectors that allow the sparse distribution between features to be minimised, thus producing more accurate results. The results of this study show an improvement in classification to 61%.

Keywords: Text classification, Multi-label, Feature selection, Weighted Information Gain, Multinomial Naïve Bayes

1.0 INTRODUCTION

The birth of electronical text processing systems such as social media platforms, instant messaging applications, online medical report repositories etc., has generated an abundance of data that ultimately challenges users to rummage through millions of documents to search for information that is most relevant to them. Therefore, in recent years, many studies have looked into finding methods to automatically classifying such information in a more organised manner [1, 2, 3] for ease of retrieval.

Text classification is a vital research area, with huge companies such as Amazon, eBay, and ImDB, looking to expand their business based on customer feedback [4]. Extracting such information is crucial as it allows business entities or policymakers [5-7] etc., to gain an insight into what drives people and how they are able to use this information to make better decisions. Similarly, the medical domain has started to recognise the impact of extracting such information from social media platforms to monitor patients' health wellbeing [8].

According to the International Diabetes Federation (IDF), diabetes is now being recognised as one of the largest global health emergencies. It is a severe illness that occurs when an individual cannot produce enough insulin or cannot use insulin and is detected by finding elevated glucose levels in the bloodstream. The IDF estimates more than a half-million children globally aged between three and fourteen years of age who are currently living with Type 1 diabetes (inability of the body to produce insulin). Four hundred fifteen million adults are already undergoing treatment for Type 1 diabetes, and an estimated 318 million adults are suffering from glucose tolerance impairment, which leaves them at a higher risk of developing the disease eventually.

The cost of medical services and the need to look for alternative medication have caused patients and caregivers to turn to online health groups to seek treatment and advice [9]. Since its emergence in the 1990s, online communities have grown from chatrooms to listservs, message boards, newsgroups, web pages, and social media sites. Studies have indicated that such groups act as priceless information sharing hubs ranging from symptoms, diagnosis discussions, and drug reactions. Communications within these groups provide insight into physicians, financial expenses, hospitalisation experiences, and daily living activities [10, 11].

Nevertheless, the abundance of available information on such platforms is not organised to make it easier for patients or caretakers to look out for relevant information [11-13]. The dissemination of information and overwhelming chat responses sometimes causes information to get lost within those platforms [10, 14]. Therefore, information must be automatically classified to eases users to search for information that would be of interest to them. This paper contribution comes in two forms: a) proposes a multi-tier framework that organises information in a more conducive manner, b) proposes to use weighted information gain feature selection method and co-trained Multinomial Naïve Bayes classifier with string vectors conversion to improve multi-label classification process.

The rest of this paper is organised in the following manner. Section 2 discusses the literature. The methodology is explained in Section 3, followed by results and discussion in Section 4. Our conclusion is in Section 5.

2.0 LITERATURE REVIEW

Diabetes has been labelled as one of the largest health emergencies that have been long overlooked. According to the 2017 World Health Organization report, more than half a million children under the age of 14 are estimated to be fighting Type 1 diabetes. Also, 415 million people worldwide are currently undergoing diabetes treatment, with 318 million more being traced to glucose resistance, potentially placing them at high risk for future diabetes. These estimates are projected to rise to 642 million individuals by 2040.

The increasing cost of medical services culminated by the desire to seek alternative treatments has directed patients and caretakers to engage with like-minded people over online health communities [9]. Research has suggested that online health communities serve as an invaluable communication and information exchange warehouse where users can interact and connect with people to discuss diagnosis, treatment options, drug side effects, etc. [15, 16, 17]. Facebook is an example of one such site. Compared to YouTube, Instagram, and Twitter, Facebook dominates the social landscape with an enormous amount of traffic reported, according to the Global Web Index Flagship 2017 survey. On average, 70 percent of the world's population has reported logging on to Facebook daily, with more than 43 percent doing so daily several times a day. Therefore, the corpus used for this research was extracted from diabetes dedicated pages on Facebook.

This paper proposes a multi-tier framework using weighted information gain feature selection and co-trained Multinomial Naïve Bayes classifier with string vectors conversion to improve the multi-label classification process. Hence, the literature of this paper will look into past studies adopting a multi-tier framework as well as studies looking to improve the classification.

2.1 Multi-Tier Classification

The availability of many textual online data has made it particularly necessary for these documents to be arranged hierarchically for better management. Research in the automatic classification of documents into pre-labelled classes has shown that it is essential to organise such data before classifying them [18]. Baqapuri et al. [19] found that the quality of classification is inversely proportional to the available data's scalability and the number of categories in which the data needs to be categorised. In other words, the time it takes for classification will suffer as the dataset increases. Nevertheless, this can be restricted by implementing a hierarchical classification that organises all categories into a tree-like structure and trains the classifier at every vertex of the hierarchy [18].

Moh et al. [4] categorised emotions for a movie review dataset using a multi-tiered system. Their classifier detected positive, negative, and neutral feedback in the first tier, followed by feedback's polarity classification. Li et al. [20] used a hierarchical filtering system where the filtering system gradually reduced online news articles' dataset hierarchically concerning contextual polarity and frequent document words. Du et al. [18] used a new weighting term that quantifies information extracted in probability distribution changes compared to TF-IDF's traditional weights and found the classifier could better classify the dataset.

With regard to the research work conducted in the past using a multi-tier classification system, only a single aspect has been attempted. The proposed framework for this research classifies data extracted from social media into two tiers using two separate elements allowing information to be organised in a more readable manner, thus promoting better classification. The following sub-section will look to discuss the tiers within the proposed framework.

2.2 Multi-Label Classification

Kanj et al. [21] defined multi-label classification as a task in which an instance can be linked to multiple classes. For example, a movie genre can be romantic (Label 1) and comedy (Label 2). Similarly, the corpus that is used within this study is also considered multi-label as the posts can be classified under more than one label. For example:

"I have been on *Metformin* for a week now and my *nausea* is not improving."
The above will be classified as treatment (*Metformin*) and Symptom (*nausea*).

The proposed framework in this study comprises of two tiers (type and purpose) where type looks to classifying the type of diabetes being referred to while purpose looks into the main topic of discussions within the extracted post (labels). Therefore, the following sub-sections will discuss past studies that have been carried out within the individual sub-sections. Table 1 showcases the literature review comparison that has been discussed in section 2.1 and 2.2.

Table 1: Comparison of the related works

Classification	Proposed Solution	Key Methods	Targeted Dataset
Multi-Tier	Baqapuri et al. [19]	Hierarchical two-level statistical classifier	Short messages on microblogging platforms.
	Moh et al. [4]	Polarity classification	Movie reviews.
	Li et al. [20]	Hierarchical filtering	Text mining from online news.
	Du et al [18]	Relaxed strategy (new weighting term based on Least Information Theory)	Newswire stories by Reuters.
Multi-Label	Kanj et al [21]	Multi-label classification as a task	Emotions, yeasts, and medical dataset.
	Ours	Weighted information gain feature selection and co-trained multinomial Naïve Bayes	Corpus extracted (related to diabetes) from dedicated Facebook pages.

2.2.1 Type Classification

Type classification in the context of this paper looks to classify posts extracted from Facebook into one of the three types of diabetes (Type1, Type2, and Gestational diabetes). International Diabetes Federation (IDF) defines Type 1 diabetes as insulin-dependent diabetes that often develops during childhood or adolescence due to insufficient insulin. The typical symptoms include excessive thirst, frequent urination, weight loss etc. Type 2 diabetes occurs when there is insufficient insulin in the human body. The symptoms of Type 2 diabetes are similar to Type 1 however the treatment options are different. Gestational diabetes occurs only during pregnancy as the placenta releases hormones that affects the mother's sugar levels.

With clinical explanation and how to differentiate one type from another, previous research on diabetes has been more medically prone [22]. El-Sappagh and Ali [23] performed a study on the ontological aspects of classifying diabetes, but ontology was only prepared for Type 1. The distinct classification for Type 2 and gestational diabetes was not available when conducting this study. This study included creating a lexicon dictionary that would cater to Type 1 diabetes and Type 2 and gestational diabetes as well.

2.2.2 Purpose Classification

Mohammad et al. [24] described purpose as the human intent upon which a tweet or post has been published. This research seeks to classify the posts extracted according to the reason(s) behind the message with the same description. For example, in the context of diabetes, the aim may be to seek information on treatment or therapy, to share a diabetic-friendly recipe, or to seek help with symptoms, etc. In other words, the purpose classification within this research will be used in the same context as finding a class for labels to be classified into.

Past literature looking into multi-label classification has adopted neural networks framework [25-28]. Although the results produced by neural networks are encouraging, it is expensive to conduct extensive training data and computational resources. Kanj et al. [21] found that current supervised learning algorithms could not correctly classify labels due to the wrong vector inputs, thus proposing an algorithm that would edit the training data to ensure the vectors assigned were not null. This is specifically useful in cases where there are too many unlabelled data available. In a study by Lee [29], a fine-grained weighting system was used as a feature selection mechanism. However, these resulted in large dimensionality issues that produced better classification in the shorter text than longer ones. This research adopted a weighted feature selection method combined with a co-trained multinomial classifier, which has been shown to improve classification by 20%.

3.0 PROPOSED FEATURE FUSION METHOD

The methodology adopted within this study is presented in this section, which includes data collection and preparation (data cleaning and pre-processing). The proposed framework is also discussed, where each tier is explained in depth within a sub-section of its own. The evaluation setup used to evaluate the proposed framework will also be introduced in this section.

3.1 Data Collection

The corpus used within this study was extracted from three diabetes-related groups on Facebook (not specified due to confidentiality reasons) established and running since 2014, with an average of 42 posts per day. The data collection duration was six months (July 2016 to January 2017) using Facebook Graph API. Before data clean-up and pre-processing, 28,048 posts and comments had to be 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 for pre-processing. This step was carried out to remove non-textual data that will compromise the classification process [30, 31].

3.2 Data Cleaning and Pre-processing

The pre-processing step helps remove negligible (noisy) data from social media text that would otherwise disrupt the classification process [32]. Elements removed include emoticons, hashtags, and URLs, non-textual posts and comments such as photos, videos etc., posts and comments that were fewer than three words, non-English text, and texts containing more than five misspelt words, among others.

Words that have been incorrectly spelt due to human error, such as sometimes spelt as sumtimes or orally spelt as orraly, were considered posts and comments containing misspelt words. For spellcheck purposes, the Wordnik dictionary was used to correct the spelling mistakes of those posts and comments that contained less than five errors. The cleaning process resulted in a total count of 21,082 raw data to work with.

A total of 6000 posts (2,000 per type of diabetes) was randomly selected from human annotation, which will ultimately train the proposed model. These 6 000 posts then went through the standard pre-processing of POS tagging, tokenisation, stop word removal, and stemming [32].

3.3 Proposed Multi-tier Classification Framework

The proposed classification framework comprises of two tiers; type and purpose (Fig. 1). Each tier will be discussed separately in the sub-sections ahead.

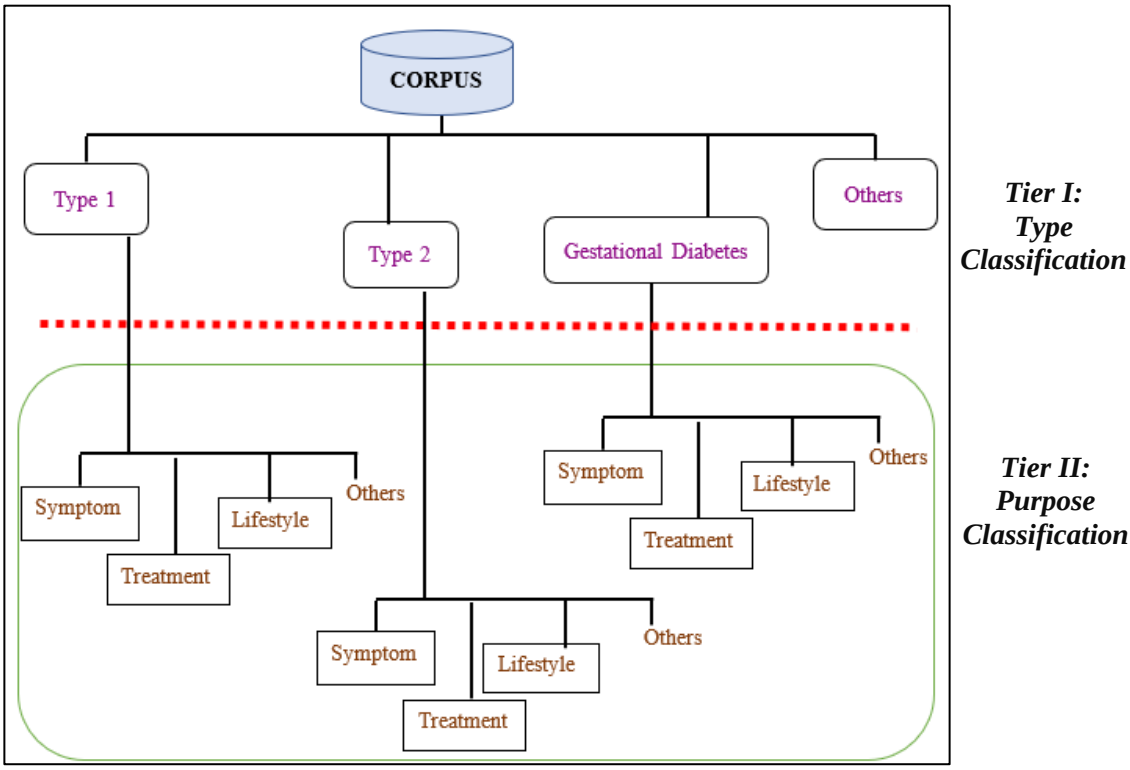

Fig. 1: Proposed Multi-tier Classification Framework

3.3.1 Type Classification

The corpus in this tier was classified into either Type1, Type2, or Gestational diabetes (Type3). For corpus text that could not be classified into any one category, the posts were then classified as Others. A manual lexicon catering to type classification was prepared for this tier. The keywords extracted from the corpus were used in the creation of this dictionary and the incorporation of a Type 1 ontology database by El-Sappagh and Ali [23]. The lexicon dictionary created caters to all three types of diabetes. In this tier, the Naïve Bayes classifier was used for type classification. The initial corpus was 6,000 posts for classification; after which only 4, 889 posts moved on to the next tier of purpose classification (Type 1= 1, 667 posts, Type 2 = 1, 590 posts, Gestational diabetes = 1, 632 posts).

3.3.2 Purpose Classification

Posts and comments are classified according to Symptom (symptoms of each type example, frequent urination, extreme fatigue etc.), Treatment (modern, traditional medicines, home remedies, clinical trials discussion etc.) and Lifestyle (exercise and diet options which include recipes shared) within this tier. Like the above tier, text that could not be classified to any one category was then classified as Others.

Since this was a multi-label classification problem where one label can overlap with another; for example, discussing the symptoms of a drug used for treating a type of diabetes could be categorised both as Treatment and Symptom, the co-training Multinomial Naïve Bayes used by Lee [29] was adapted using weighted Mutual Information gain as a feature selection. The feature selection was modified to focus on the posts that were wrongfully classified and re-assign the weights accordingly. To improve classification, vectors were converted to string vectors, which allowed the sparse distribution within features to be reduced, thus allowing for better classification [33]. Fig. 2 shows the weighted information gain feature selection that was fed into the dependent classifier of the co-training Multinomial Naïve Bayes classifier.

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

$S \leftarrow \emptyset$;
$W \leftarrow 1$ {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}
  $S \leftarrow S \cup F_{max}$ {Add feature to subset}
  $F \leftarrow F \setminus F_{max}$ {Remove feature from candidate set}
  Classifier $\leftarrow$ Train DC Classifier (X, S, Y)
  $Y' \leftarrow$ ApplyClassifier (Classifier, X)
  $W \leftarrow |Y - Y'|$ {Residual of each sample is new weight}
  CheckStoppingCriterion ()
end while

Fig. 2: Feature Selection Pseudocode

3.4 Experimental Setup

Standard evaluation metrics were used to evaluate the proposed framework, along with comparisons with benchmark models. The evaluation of each tier is done separately as each tier adopts a different technique. Metrics used include 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 2) [34] and multi-label classification (Fig. 3) [35].

Table 2: Confusion Matrix for binary classification [34]

		Predicted	
		False	True
Actual	False	TN	FP
	True	FN	TP

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

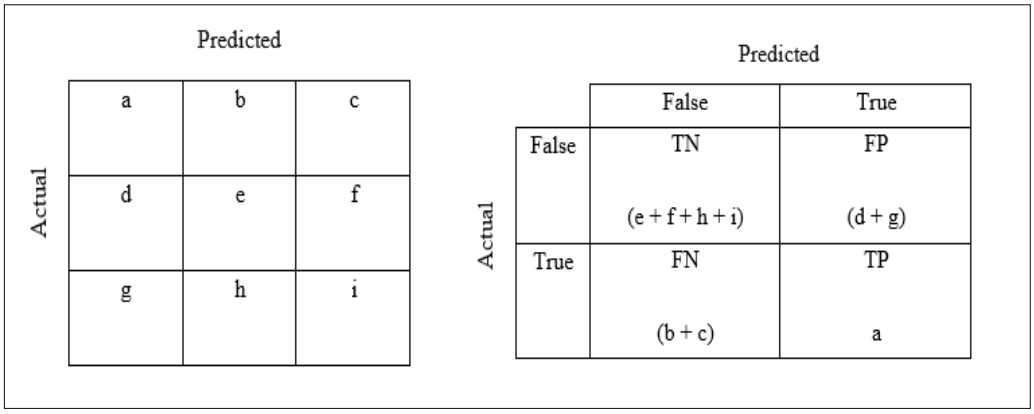

Fig. 3: Multi-Label confusion matrix [35]

Table 2 and Fig. 3 is generated from the following four measures [34, 35]:

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

The evaluation was calculated using ten-fold cross-validation. The equations for accuracy, F1-Score, and Area Under Curve (AUC), respectively, were adopted from Idrees et al. [34], Ruuska et al. [35], and Anand and Naorem [36].

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

$$F1 - Score = \frac{2\ x\ precision\ x\ recall}{precision + recall} \quad (2)$$

$$AUC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \quad (3)$$

Apart from the above, the purpose classification tier's evaluation used three multi-label performance measures [37], namely Hamming Loss, 0/1 Loss, and accuracy. Hamming Loss equation (4) treats each label as a distinct binary evaluation while the 0/1 Loss (5) measure states any predicted label must match the true set of labels (c) exactly

$$Hamming\ Loss = 1 - \frac{1}{NL}\sum_{i=1}^{N}\sum_{l=1}^{L} 1(c_l^i = c_l^{\wedge i}) \quad (4)$$

$$0/1\ Loss = 1 - \tfrac{1}{NL}\textstyle\sum_{i=1}^{N} 1(c^i = c^{\wedge i}) \quad (5)$$

The accuracy measure for multi-label classification (6) used for evaluation purpose classification tier has been used as the standard evaluation technique in past multi-label classification problems [29, 38-40]

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{c^i \wedge c^{\wedge i}}{c^i\ \mathrm{v}\ c^{\wedge i}} \quad (6)$$

The experiments were conducted using several benchmark models. The models used were as follows:

- $M_{Reichert}$: classification using Reichert et al. [41] benchmark model
- M_{Salas}: classification using Salas-Zárate et al. [42] benchmark model
- M_{TP-t} : classification using proposed classification framework

4.0 RESULT & DISCUSSION

The results of each tier within the proposed framework will be discussed in this section. Each sub-section will look to display and discuss the results separately for ease of understanding.

4.1 Tier 1: Type Classification

For benchmarking purpose, two models were used ($M_{Reichert}$ and M_{Salas}). The comparison with benchmarking models was only done for Type 1 and Type 2 diabetes. As M_{Salas} comprises of Type 1 tweets while $M_{Reichert}$ consists of Type 2 online forum posts. Table 3 shows the classification results for Type 2 ($M_{Reichert}$) is more encouraging compared to Type 1 (M_{Salas}). When analysing the dataset, it was found that the length of the forum text was almost as long as posts extracted from Facebook and contained almost the same lingo and jargon. On the other hand, Tweets require a different form of data cleaning and pre-processing; hence, most of the data was lost in the cleaning process as it was deemed misspelt words [43, 44].

Table 3: 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_{TP-t} (T1)	**0.77**	**0.76**	**0.77**
M_{TP-t} (T2)	0.69	0.69	0.69
M_{TP-t} (T3)	0.76	0.75	0.76

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

The proposed framework was able to classify Type 1 more accurately than Type 2 and Type 3 concerning the classification results of each type of diabetes (Table 3). This may relate to the lexicon used within this tier, an extension of the type 1 diabetes ontology used by El-Sappagh and Ali [23]. There are more words in the lexicon dictionary that cater for type 1 diabetes, and so the ability to match more words for type 1 diabetes may have contributed to better results in the classification. Similarly, the classification scores for gestational diabetes (i.e., Type 3) also proved better than Type 2. Again, this is probably because of matching the keyword between the dictionary of the lexicon and the words used within the dataset.

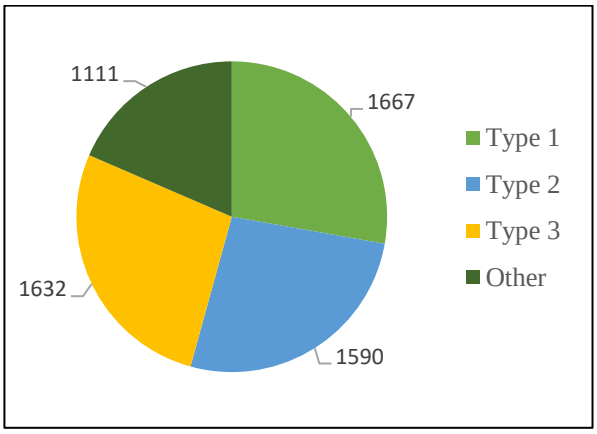

Fig. 4: Posts classified per diabetes type

Fig. 4 shows the number of posts classified per type. Posts that could not be classified into any of the three types of diabetes were categorised as Other (N = 1, 111). It can be seen from the figure that most posts belong to type 1, followed by gestational diabetes (Type 3) and Type 2. The following sub-section would discuss the outcomes of the classification of purpose where the sum of data carried from this tier to the next was 4,889 posts.

4.2 Tier 2: Purpose Classification

A feature selection experiment was conducted to determine the type of feature selection approach that would better fit this classification level. Six of the most commonly used feature selection techniques [45]; 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 text classifiers [46]; Naïve Bayes, Support Vector Machine and Logistic Regression. F1-Score results of the experiments are as displayed against the number of features using different classifiers (Fig. 5, Fig. 6, Fig. 7), while Table 4 showcases the best feature selection technique based on the number of features (per hundred).

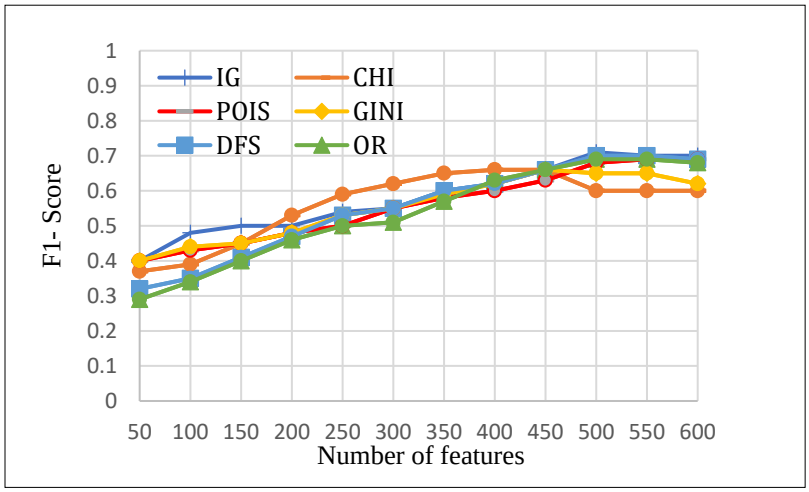

Fig. 5: F1 Score Using Naive Bayes

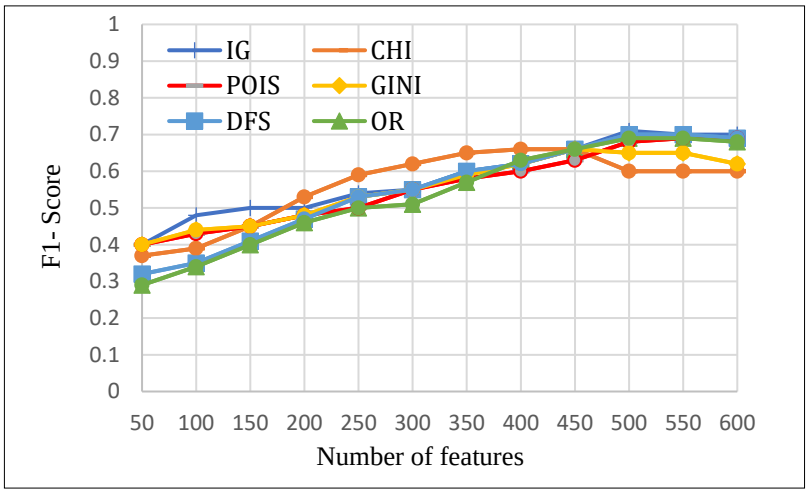

Fig. 6: F1 Score Using Support Vector Machine

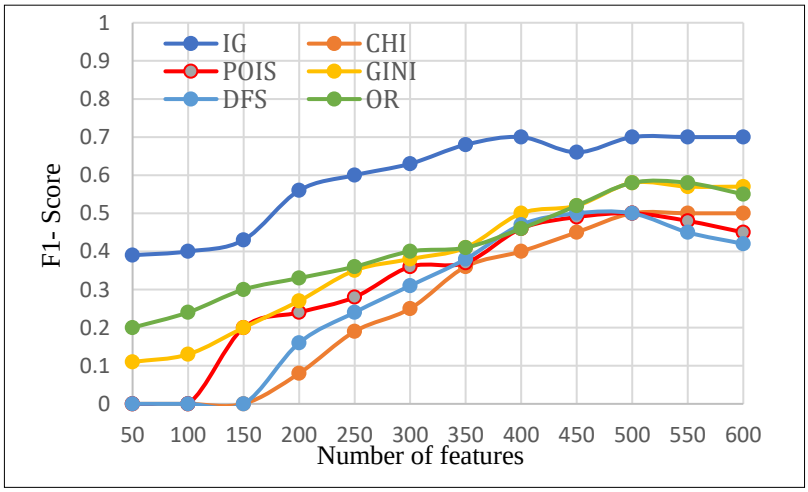

Fig. 7: F1 Score Using Logistic Regression

Table 4: Feature Selection Technique Results

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*

The results (Fig. 5, 6, and 7) show that the optimum results were achieved when features were set to 500. Furthermore, amongst the three classifiers used in this experiment, Naïve Bayes proved to show promising results (Fig. 5). Both Support Vector Machine (Fig. 6) and Logistic Regression (Fig. 7) produced zero F1-Scores for features between 50 and 150, yet Naïve Bayes worst F1-Score was 0.29 despite the low number of features set (50 features). Literature has also accepted that when it comes to multi-label or multi-class classifications, Naïve Bayes works best [29, 39]. Therefore, this study adopted the Information Gain feature selection technique and the Naïve Bayes classification algorithm for this classification level.

In the earlier stages of experiments, ten labels were identified. However, it became apparent that certain labels overlapped each other and caused the F1-Score of the classification to suffer. Therefore, some labels had to be combined to counter this problem. The results of the trial experiments are shown in Fig. 8.

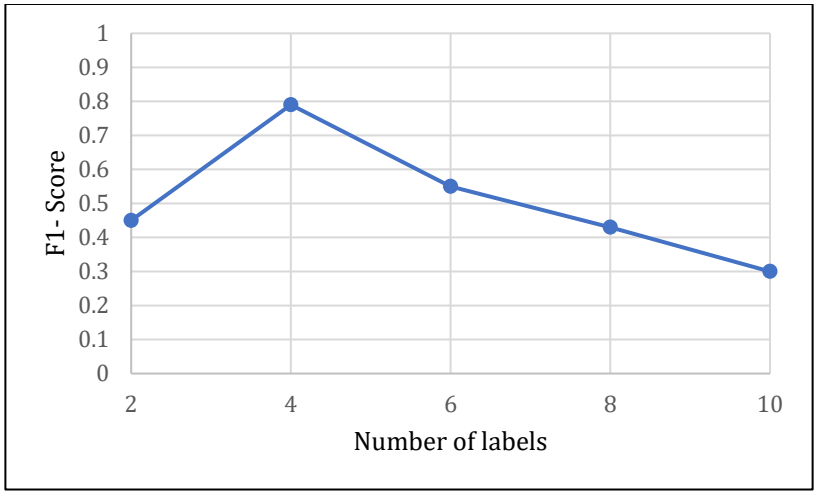

Fig. 8: F1-Score with respect to number of labels

An increase in the number of labels resulted in a decrease in F1-Score (Fig. 8), suggesting that the classifier's performance was deteriorating. This was due to insufficient labelled data available for training. Hence, similar labels under a more general label were merged, which improved the F1-Score considerably. Instead of classifying Metformin as modern medicine, for example, and herbal mixture as traditional medicine, both (Metformin and herbal mixture) were categorised under the Treatment mark instead. Consequently, it was realised that the problem at hand was a multi-label classification problem. For example,

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

Herbal tea can be branded as therapy and lifestyle improvements from the above sample, making it a multi-label problem. To resolve this, the literature suggests the use of Multinomial Naïve Bayes (MNB) [29, 47, 48]. Nevertheless, experiments conducted using MNB alone did not produce favourable FI-Scores and AUC (Table 5). Therefore, other experiments were carried out using co-training with weighted Information Gain feature selection and string vectors, which eventually improved results.

Table 5: Results 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 operates by defining characteristics and labels, where it identifies and classifies words such as *Metformin daily* as a bigram function under the label of treatment. The Symptom label provided the highest F1-Score and AUC, although the highest number of data available for purpose classification belonged to the Treatment label (Fig. 9). Symptom had the most clearly defined boundaries, which allowed the algorithm to identify it easily. For example:

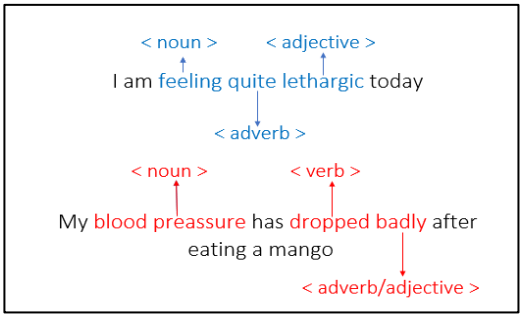

Fig. 9: Sample Text

During cross-checking for referenced nouns, the classifier preferred adjectives that either preceded or succeeded adverbs (Fig. 9) within the post before categorising it as Symptom. The weighted information gain feature selection technique re-adjusted weights for incorrectly classified features, and the F1-Score and AUC for Symptom provided better scores because the features were more distinct for the Symptom label than the other two. Similarly, it was a little trickier to define a post that belongs distinctly to the Lifestyle and Treatment choice as several posts could either belong to either one. For example:

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

The above indicates a shift of lifestyle to a diet, but it can also be viewed as a cure for home remedies. One hundred inconsistent posts were sent for annotation to assess further whether a potential trend could help the algorithm differentiate between the label Treatment and Lifestyle. The Krippendorff alpha was measured at 0.62, which is not acknowledged as the right consensus rate between annotators [49]. However, it is speculated that the F1-Score and AUC could perhaps be boosted if the volume of data available for training the algorithm was extended.

Hamming Loss, 0/1 Loss, and Accuracy are accepted as the standard evaluation measures in assessing multi-label classification algorithms, as mentioned in the methodology section [29, 38-40]. Hamming Loss explores the individual labels that were wrongly expected, while 0/1 Loss looks at the entire set of labels. Therefore, if it does not fit the true set of labels, the entire set of labels in a sample post will be deemed incorrect. Table 6 depicts the results of each label based on these metrics.

Table 6: Evaluation Metrics Results

	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 in this classification rank used individual characteristics as input into the dependency mark algorithm. The Hamming Loss results are thus reported as the lowest, as it is measured for label-based assessment. Conversely, the 0/1 loss metric is optimised for label-based assessment, and the co-training algorithm manipulates whole labels for classification purposes, so the 0/1 loss results are reported as the highest of the three.

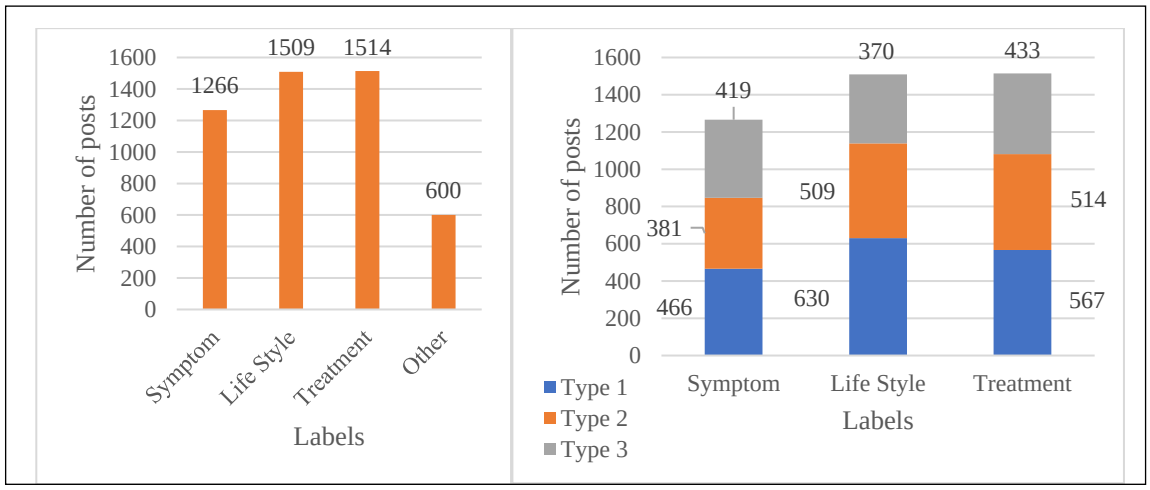

Fig. 10: Posts Classified by Purpose

Fig. 10 is a graphical depiction of the breakdown of data within the purpose labels. It should be noted that the purpose of the classifier has been the most capable of classifying Type 1 posts for the Life Style label. This was an obvious choice because several recipes and workout choices were categorised as Life Style in the training results. The algorithm proceeded to re-assign weights to the wrongly defined labels with the aid of the weighted information gain feature selection, thus improving its accuracy. The Symptom label contains the least amount of details because fewer symptom posts were included in the context of the dataset, but much more towards recovery options and improvements that can be made to everyday life that may enhance patients' quality of life.

This tier's final analysis was focused on the output of the classifier for each type. The F1-Score obtained for each reason (label) per form of diabetes is shown in Table 7. Life Style scored the highest for Type 2 diabetes, while treatment registered the lowest F1-Score, even for Type 2. This was due to the structure of the data from training fed into the algorithm. Since type 2 diabetes is a far more regulated form of diabetes, this column's advice is therefore far more linked to healthier snacks and diet choices followed by home remedies that may help postpone the disease's effects. Metformin is the most prominent medication known for type 2, although it is difficult to assess from the text itself whether the recommended medication is intended for type 1 or type 2, which explains the low F1-Score for Type 2 Treatment.

Table 7: 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

5.0 CONCLUSION, LIMITATIONS & FUTURE WORK

The available literature analysis notes that many users turned to online health support groups to seek help from those who were also fighting the same condition or merely seeking advice from others on the available treatment options [14-16]. However, it was not easy to locate the correct information considering the vast data availability [15, 17]. Therefore, this paper's goal was to automatically identify posts extracted from Facebook to enhance the process of classification, thereby providing users with information that better fits their needs. In politics [24] and within the field of goods and services [5, 28, 36], recent research aimed at classifying opinion, emotion, and purpose has accomplished so. Nevertheless, in order to create better classifications, the methods used within these studies may be improved.

Using assessment metrics (F1-Score, the area under curve and accuracy) and benchmark dataset comparisons, each tier was measured separately within the proposed system. In contrast to the benchmark research, the proposed system was found to produce more detailed classifications, thus validating the techniques adopted. Notably, for type classification, the proposed framework provided 77 percent of F1-Score than the benchmark (i.e., 70 percent). Similarly, the addition of co-training along with a weighted feature selection technique and string vector conversion boosted the F1-Score from 38% to 61%.

Nonetheless, for posts in other languages, the dataset used to train the algorithm consisted of posts that were only in English, and so the framework would also not be able to perform well. We plan to extend our system to support other common languages such as Arabic and Spanish in our future work. In the context of appealing to irony and sarcasm, the next constraint arrives. Both irony and sarcasm are described as a negative feeling disguised as a positive feeling [50, 51]. Without evaluating for sarcasm or irony, this analysis took each post at face value, which may lead to a different sentiment score and emotion. This is because a satirical post might come across as a joy, close to emotion, but it could be in spite in indirect ways. To improve the classifications of feeling and emotion, we will consider sarcasm and irony in our future works.

6.0 ACKNOWLEDGEMENT

The authors would like to thank and acknowledge the support provided by the University of Malaya, under research grant reference number: UMRG RP059C 17SBS.

REFERENCES

[1] S. Marcos-Pablos and F. J. García-Peñalvo, Information retrieval methodology for aiding scientific database search, Soft Computing, vol. 24, no. 8, pp. 5551-5560, 2020.

[2] Stein, R. A., Jaques, P. A. and Valiati, J. F. An analysis of hierarchical text classification using word embeddings. Information Sciences, 471 (2019), 216-232.

[3] Liu, J., Chang, W.-C., Wu, Y. and Yang, Y. Deep learning for extreme multi-label text classification. City, 2017.

[4] Moh, M., Gajjala, A., Gangireddy, S. C. R. and Moh, T. S. On Multi-tier Sentiment Analysis Using Supervised Machine Learning. City, 2015.

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

[6] Alashri, S., Srivatsav Kandala, S., Bajaj, V., Parriott, E., Awazu, Y. and C Desouza, K. The 2016 US Presidential Election on Facebook: An Exploratory Analysis of Sentiments. City, 2018.

[7] Sandoval-Almazan, R. and Valle-Cruz, D. Facebook impact and sentiment analysis on political campaigns. ACM, City, 2018.

[8] Benetoli, A., Chen, T. and Aslani, P. How patients' use of social media impacts their interactions with healthcare professionals. Patient education and counseling, 101, 3 (2018), 439-444.

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

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

[11] Zhang, Y., He, D. and Sang, Y. Facebook as a platform for health information and communication: a case study of a diabetes group. Journal of medical systems, 37, 3 (2013), 9942.

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

[13] Willis, E. and Royne, M. B. Online health communities and chronic disease self-management. Health communication, 32, 3 (2017), 269-278.

[14] Park, H., Reber, B. H. and Chon, M.-G. Tweeting as health communication: health organisations' use of twitter for health promotion and public engagement. Journal of health communication, 21, 2 (2016), 188-198.

[15] Sharma, M., Yadav, K., Yadav, N. and Ferdinand, K. C. Zika virus pandemic—analysis of Facebook as a social media health information platform. American journal of infection control, 45, 3 (2017), 301-302.

[16] Aref-Adib, G., O'Hanlon, P., Fullarton, K., Morant, N., Sommerlad, A., Johnson, S. and Osborn, D. A qualitative study of online mental health information seeking behaviour by those with psychosis. BMC psychiatry, 16, 1 (2016), 232.

[17] K. Obamiro, S. West, and S. Lee, Like, comment, tag, share: Facebook interactions in health research, International Journal of Medical Informatics, vol. 137, p. 104097, 2020.

[18] Du, Y., Liu, J., Ke, W. and Gong, X. Hierarchy construction and text classification based on the relaxation strategy and least information model. Expert Systems with Applications, 100 (2018), 157-164.

[19] Baqapuri, A. I., Saleh, S., Ilyas, M. U., Khan, M. M. and Qamar, A. M. Sentiment classification of tweets using hierarchical classification. City, 2016.

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

[21] Kanj, S., Abdallah, F., Denoeux, T. and Tout, K. Editing training data for multi-label classification with the k-nearest neighbor rule. Pattern Analysis and Applications, 19, 1 (2016), 145-161.

[22] Association, A. D. 2. Classification and diagnosis of diabetes. Diabetes care, 39, Supplement 1 (2016), S13-S22.

[23] El-Sappagh, S. and Ali, F. DDO: a diabetes mellitus diagnosis ontology. SpringerOpen, City, 2016.

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

[25] Baker, S. and Korhonen, A.-L. Initialising neural networks for hierarchical multi-label text classification. Association for Computational Linguistics, City, 2017.

[26] Gargiulo, F., Silvestri, S., Ciampi, M. and De Pietro, G. Deep neural network for hierarchical extreme multi-label text classification. Applied Soft Computing, 79 (2019), 125-138.

[27] Ive, J., Gkotsis, G., Dutta, R., Stewart, R. and Velupillai, S. Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health. City, 2018.

[28] Poria, S., Cambria, E. and Gelbukh, A. Aspect extraction for opinion mining with a deep convolutional neural network. Knowledge-Based Systems, 108 (9/15/ 2016), 42-49.

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

[30] Toujani, R. and Akaichi, J. Fuzzy sentiment classification in social network facebook'statuses mining. IEEE, City, 2017.

[31] Akaichi, J. Social networks' Facebook'statutes updates mining for sentiment classification. IEEE, City, 2013.

[32] Singh, T. and Kumari, M. Role of Text Pre-processing in Twitter Sentiment Analysis. Procedia Computer Science, 89 (// 2016), 549-554.

[33] Jo, T. *Improving K Nearest Neighbor into String Vector Version for Text Categorisation*. IEEE, City, 2019.

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

[35] Ruuska, S., Hämäläinen, W., Kajava, S., Mughal, M., Matilainen, P. and Mononen, J. Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle. *Behavioural processes*, 148 (2018), 56-62.

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

[37] Read, J., Pfahringer, B., Holmes, G. and Frank, E. Classifier chains for multi-label classification. *Machine learning*, 85, 3 (2011), 333

[38] Elghazel, H., Aussem, A., Gharroudi, O. and Saadaoui, W. Ensemble multi-label text categorisation based on rotation forest and latent semantic indexing. Expert Systems with Applications, 57 (9/15/ 2016), 1-11.

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

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

[41] Reichert, J.-R., Kristensen, K. L., Mukkamala, R. R. and Vatrapu, R. *A supervised machine learning study of online discussion forums about type-2 diabetes*. IEEE, City, 2017.

[42] Salas-Zárate, M. d. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodríguez-García, M. Á. and Valencia-García, R. Sentiment analysis on tweets about diabetes: an aspect-level approach. *Computational and mathematical methods in medicine*, 2017 (2017).

[43] Chandra Pandey, A., Singh Rajpoot, D. and Saraswat, M. Twitter sentiment analysis using hybrid cuckoo search method. *Information Processing & Management*, 53, 4 (7// 2017), 764-779.

[44] Tellez, E. S., Miranda-Jiménez, S., Graff, M., Moctezuma, D., Siordia, O. S. and Villaseñor, E. A. A case study of Spanish text transformations for twitter sentiment analysis. *Expert Systems with Applications*, 81 (9/15/ 2017), 457-471.

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

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

[47] Aldoğan, D. and Yaslan, Y. A comparison study on active learning integrated ensemble approaches in sentiment analysis. *Computers & Electrical Engineering*, 57 (1// 2017), 311-323.

[48] Lee, J. and Kim, D.-W. SCLS: Multi-label feature selection based on scalable criterion for large label set. *Pattern Recognition*, 66 (6// 2017), 342-352.

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

[50] Mukherjee, S. and Bala, P. K. Sarcasm detection in microblogs using Naïve Bayes and fuzzy clustering. *Technology in Society*, 48 (2// 2017), 19-27.

[51] Ravi, K. and Ravi, V. A novel automatic satire and irony detection using ensembled feature selection and data mining. *Knowledge-Based Systems*, 120 (3/15/ 2017), 15-33

[52] *Diabetes Facts & Figures.* (2020, December 12). International Diabetes Federation. https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html