European Journal of Computer Science and Information Technology
Vol.8, No.5, pp.46-66, November 2020
Print ISSN: 2054-0957 (Print),
Online ISSN: 2054-0965 (Online)

EMPIRICAL STUDY OF FEATURES AND UNSUPERVISED SENTIMENT ANALYSIS TECHNIQUES FOR DEPRESSION DETECTION IN SOCIAL MEDIA

Shahad Ayedh Alharthi
Master of Science in Computing (Data Analytics)
Riyadh – Saudi Arabia
Dublin City University
shahad.alharthi222@gmail.com

ABSTRACT: *This study provides an empirical evaluation of diverse traditional learning, deep learning, and unsupervised techniques based on diverse sets of features for the problem of depression detection among Twitter and Reddit users. The main objective of this study is to investigate the most appropriate features, document representations, and text classifiers for the significant problem of depression detection on social media microblogs, such as tweets, as well as macroblogs, such as posts on Reddit. The study's investigation will concentrate on the linguistic characteristics, blogging behavior, and topics for features, multi-word, and word embeddings for document representation as well as on unsupervised learning for text clustering. This study will select the best approaches in the literature as baselines to practically examine them on the depressive and non-depressive dataset of blogs designed for this work. The study's integrations and ensembles of the selected baselines will be experimented as well to recommend a design for an effective social media blog classifier based on unsupervised learning and WE document representation. The study concluded that the experiments proved that a stacking ensemble of Adam Deep Learning with SOM clustering followed by Agglomerative Hierarchical clustering with topic features and pre-trained word2vec embeddings achieved an accuracy more than 92% on Twitter and Reddit depression analysis datasets.*

KEYWORDS: empirical study, analysis, depression, social media, data analytics, computing.

INTRODUCTION

According to the World Health Organization (WHO) [1], more than 300 million people all over the world suffer from depression, the mental disorder causing disability in an individual's life, and gradually developing into commit suicide. Usually, depressed individuals do not know about their situation although they express the symptoms through noticeable behavior. Depressed individuals, for instance, tend to keep silent, fatigue, and socially isolated in real-life while on social media platforms, they tend to express their feelings and thoughts a bit more actively. Previous research and clinical methods for depression diagnosis already proved the direct effect of this mental disorder on the language of the patient. To illustrate, using the Linguistic Inquiry and Word Count (LIWC) [3], it has been observed that depressed people frequently use the first person singular

pronouns, word "I", in contexts of death, sad and negative in addition to the use of verbs in the past tense and absolutist words such as “absolutely”, “completely”, “always”, and “never” [2, 4, and 12]. As social media platforms are considered huge warehouses of information about their users, the need for automatic diagnosis of depression symptoms out of text content is gaining importance.

Natural Language Processing (NLP) and Machine Learning (ML) techniques have been extensively applied for text classification problems in fields such as sociology, psychology, and sentiment analysis [6]. Depression detection for social media users can be formulated as a text classification problem where the target classes being “Depressed” and “Non-Depressed” users and the document to analyze are a set of selected posts in a user’s social media timeline [6]. Recent studies that approached depression detection have incorporated supervised and unsupervised traditional ML, Deep Learning (DL) as well as Word Embedding's (WE) techniques [7]. However, there is no agreement yet regarding which method is the best. While supervised ML and DL approaches achieved high prediction scores compared to unsupervised, their most notable concern is the unfeasibility of providing sufficient amounts of annotated training datasets [7]. Unsupervised learning, on the other hand, allows to group similar data samples into clusters without previously being trained on annotated data [5]. However, as the generation of clusters is based on unsupervised similarity measurements [5], the resulted clusters might not accurately correspond to the required classes for a particular classification problem. To illustrate, if a clustering approach is given a set of depressed and non-depressed bloggers then, without a proper feature engineering, it might cluster them based on the topics they are blogging about or on their geographical location rather on being depressed or non-depressed.

Therefore, the process of feature selection and vector representation that highlights the distinction between depressive and non-depressive text is the heart to the success of a particular text classification problem [5]. Recently, WE techniques, which represent every word with a vector of its synonyms hence the semantic similarity of the text is being captured, have been widely used for depression detection and sentiment analysis in social media blogs [8, 9]. It has been proved with experimental results that WE significantly enhanced the performance of many state-of-art supervised classification methods [7, 8, and 9]. Hence, the investigation of such an approach will be the eventual objective of this study.

Study Objectives and Structure

According to the previous introduction, the main objective of this study is to investigate the most appropriate features, document representations and text classifiers for the significant problem of depression detection on social media microblogs, such as tweets, as well as macroblogs, such as posts on Reddit. In particular, the investigation will concentrate on the linguistic characteristics, blogging behavior, and topics for features, multi-word and word embedding's for document representation as well as on unsupervised learning for text clustering. This study will begin by selecting the best approaches in the literature as baselines to examine them practically on the depressive and non-depressive dataset of blogs designed for this work. Next, integrations and ensembles of the selected

baselines will be experimented as well in order to eventually recommend a design for an effective social media blog classifier based on unsupervised learning and WE document representation.

LITERATURE REVIEW

This chapter reviews multiple methods and techniques applied for text classification tasks in the particular domains of depression detection, mental health disorders, sentiment analysis, and document clustering. The techniques can be applied in the activities of feature selection, topic modeling, document representation, and classification. The literature has been conducted based on the following methodology and criteria.

Search Methodology and Literature Selection Criteria

The searches were conducted in scientific digital libraries such as Citeseerx, Google Scholar, IEEE Xplore, NCBI, PsycNet, Research Gate, Science Direct and Semantic Scholar using the main keywords such as "depression detection", "sentiment analysis", "social media analysis", "machine learning", "deep learning" and "unsupervised learning".

Feature Selection

Most of the literature on mental disorders in social media has focused on feature selection. It was noticed that the extraction of lexical features using the LIWC dictionary, which contains more than 32 categories of different psychological and linguistic contexts, is the most adopted feature for mental disorder analysis such as [10]. LIWC feature extraction [3] reads the text, parses each word to a pre-defined category, and then counts all words inside every category. LIWC categorization proved its efficiency since it has been developed in psychology and sociology [2, 3]. In 2017, Shen G. et al. [11] have identified and extracted six feature groups related to depression to comprehensively describe each user as the following:

- **Social Network Features:** The study of [12] analyzed three datasets; depressive users, depressive tweets, and control tweets with statistical methods. It concluded that depressive users are less active in posting tweets, doing it more often in the time range from 23:00 to 6:00.
- **User Profile Features:** The profile features highlight personal information such as age, gender, etc. It was noticed that individuals having an academic degree or career are less likely to be depressed [11].
- **Emotional Feature:** According to [11], emotional terms of depressed users differ from that of other users, so the emotional features are beneficial in depression detection.
- **Part-of-Speech (POS) Tagging:** [19] defined POS as a method of splitting the sentences into words and attaching a proper tag such as noun, verb, adjective, and adverb to each word based on the POS tagging rules. This suggests that POS is preferred when unsupervised text clustering is applied.

❖ **Topic Feature:** The ideas concerned by depressed and non-depressed individuals are likely to differ [11]. For [13], topic modeling is an unsupervised text mining approach that discovers similarities in a text or a corpus that speaks about the same ideas.

Document Representation

Throughout the literature about NLP, many document representation methods have been used:

❖ **Bag-Of-Words (BoW):** The name BoW refers to the dictionary built by indexing words (or groupings of words) in all documents of the corpus. Each index in the dictionary can correspond to a uni-gram (one word) or an n-gram (sequence of n consecutive words).

❖ **Multi-Word:** Another method to reflect text semantics than uni-gram topic features could the n-grams word features (i.e. multi-word) generated by count or TFIDF vectorizer. In 2019, the work [15] investigated the LIWC, LDA, uni-gram and bi-gram BOW features with supervised ML classifiers such as Logistic Regression (LR), SVM, RF, Ada Boost and Multilayer Perceptron (MLP) to detect depression in Reddit posts.

❖ **Word Embeddings (WE):** Sometimes a single word with the same pronunciation and even the same spelling can have multiple meanings. Like the word "apple‟ is a "company" name as well as a "fruit‟. Mikolov, et al. [20] introduced Word2vec, a prediction-based embedding. Vectors in Word2vec can be obtained using two methods (both involving Neural Networks): Skip-Gram and Common Bag-of-Words (CBOW). The CBOW method predicts the probability of a word given a context.

Supervised Classification

The supervised learning methods depend on the existence of labeled training documents. In supervised machine learning, there are two data set are used. One data set known as the train data set is used to train the data on the selected classifier. In the field of depression detection, extensive use of supervised approaches were proposed recently [7].

Unsupervised Classification

It is mainly used for feature reduction tasks to minimize the size of the dataset and/or feature vectors. In comparison with unsupervised classification [5], a supervised approach, with manual labeling of the dataset, in case of social networks data analysis, potentially leads to a time-consuming effort that could be unfeasible in certain scenarios.

Evaluation Metrics

These metrics describe the evaluation metrics applied in this work where TP stands for the number of true positive predictions, TN stands for true negative predictions, FP stands for false-positive predictions, and FN stands for false-negative predictions. As suggested by [21], to select an evaluation metric to decide on the preferred model to detect depression, the following measures are suggested:

- If the emphasis is on the ability to capture and identify depressed individuals at risk to encourage them for checkup and treatment.

- If the emphasis is on making accurate statistics on depressed users then accuracy and AUC are prioritized.
- The emphasis is on computational resources and runtime, then a preferred model is that with high to moderate metrics values coupled with as low computational resources and runtime as possible.

METHODOLOGY

To address the research questions of this study, the following three experiments will be conducted:

Baselines

In this experiment, best-unsupervised baselines are selected from the literature and then each baseline will be tested on Twitter and Reddit datasets. The selected **unsupervised baselines** from the literature include the following:

1. The K+DENCLUE as proposed by work [16].
2. The SOM+Adam as proposed by work [17].
3. The AHCD as proposed by work [18]

The selected **features** from the literature are the following:

1. LIWC features as identified in work [15].
2. Social Network Features (SNF) as identified in work [11].
3. User Profile Feature (UPF) as identified in work [11].
4. Emotional Feature (EMF) as identified in work [11].
5. Topic Feature (TPC) as identified in work [13].
6. Part-of-Speech (POS) as identified in work [19].

The selected **document representations** from the literature include the following:

1. Unigrams of ssToT topics [13] with TFIDF weights as in work [14].
2. Bigrams of document words.
3. Word embeddings (pre-trained FastText).
4. Word embeddings (pre-trained Word2Vec).

Implementation of baselines will be using Gensim, sklearn, Spacy, and other machine learning libraries in Python in addition to third party code.

Baselines with Combinations of Features and Document Representation

In this experiment, for each unsupervised baseline identified in 3.1, the following combinations of features will be tested on Twitter and Reddit datasets and represented using each document representation identified in 3.1:

1. UPF.
2. SNF.
3. EMF.
4. POS.
5. TPC.

6. UPF + SNF
7. TPC + POS
8. UPF + SNF + EMF
9. TBC + POS + EMF
10. SNF + UPF + TBC + POS
11. SNF + UPF + TBC + POS + EMF

Ensembles of Baselines

In this experiment, the highest score combination of features and document representation in experiment 3.2 will be tested with two types of ensembles of unsupervised baselines identified in 3.1 on Twitter and Reddit datasets. The first ensemble method is the Voting and the second ensemble method is Stacking.

Voting

In this ensemble type, multiple baselines execute at the same time and generate predictions. The accepted prediction among them is the one that was generated by the maximum number of baselines or the average of predictions. The following voting ensembles are tested:

1. {K+DENCLUE, SOM+Adam}.
2. {K+DENCLUE, AHDC}.
3. {SOM+Adam, AHDC}.
4. {K+DENCLUE, SOM+Adam, AHDC}.

Stacking

In this ensemble type, the baselines will execute in a sequence such that each baseline (except for the first one) will add the prediction of the previous baseline to the feature vector it will process. The accepted prediction will be the one generated by the last baseline. The following stacking ensembles are tested:

1. {AHDC => SOM+Adam}
2. {AHDC => K+DENCLUE}
3. {AHDC => SOM+Adam => K+DENCLUDE}
4. {AHDC => K+DENCLUE => SOM+Adam}
5. {SOM+Adam => AHDC}
6. {SOM+Adam => K+DENCLUE}
7. {SOM+Adam => AHDC => K+DENCLUE}
8. {SOM+Adam => K+DENCLUE => AHDC}
9. {K+DENCLUE => SOM+Adam}
10. {K+DENCLUDE => AHDC}
11. {K+DENCLUDE => SOM+Adam => AHDC}
12. {K+DENCLUDE => AHDC => SOM+Adam}

Data and Analysis

The dataset collected for this work consists of the following Twitter microblogs and Reddit macroblogs:

1. One million non-depressed tweets from the Sentiment140 dataset[1].
2. 200,000 depressive tweets[2].
3. 33,000 non-depressed Twitter users[3].
4. 1,600 depressed Twitter users[4,5].
5. 1,500 non-depressive Reddit posts (under the category of CasualConversation)[6].
6. 2,800 depressive Reddit posts (under the category of SuicideWatch)[6, 7]
7. 960 depressive Reddit posts (under the category of Depression)[7]

For social media users, the tweets/posts in each timeline are scraped since one year ago and then appended together such that each document belongs to one user timeline. The pre-processing steps of the data will include removing hashtag symbols and underscore from hashtags, removing URLs and user mentions, translating acronyms and internet slangs using third party lexicons, correcting spelling errors, removing of punctuations, lemmatization and finally tokenization. Lemmatization is considered as normalization of the words by returning all words to their roots. To illustrate, words such as "goes", "going", "gone" and "went" are all converted into "go". This step will reduce the dimensionality of the feature vector and capture the similarity between documents of each class accurately.

Frequency of Words in Depressed vs. Non-Depressed Users

The first linguistic analysis of the depressive text is the count of different types of pronouns, absolute words, and negative emotional words. With Figures (1) and (2), it is noticed that the most occurring pronouns in depressive tweets are the first person singular pronouns such as "I" and "my" in addition to the second-person pronouns such as "you".

[1] https://www.kaggle.com/kazanova/sentiment140
[2] https://github.com/bonn0062/tweemotions
[3] https://github.com/goonmeet/Depression-User-Profile-Twitter
[4] https://github.com/goonmeet/Depression-Tool
[5] https://github.com/munniomer/depression-detection-using-twitter-data
[6] https://github.com/Allenfp/DepressionDetectionNLP
[7] https://github.com/thelee81/GA-Project-3

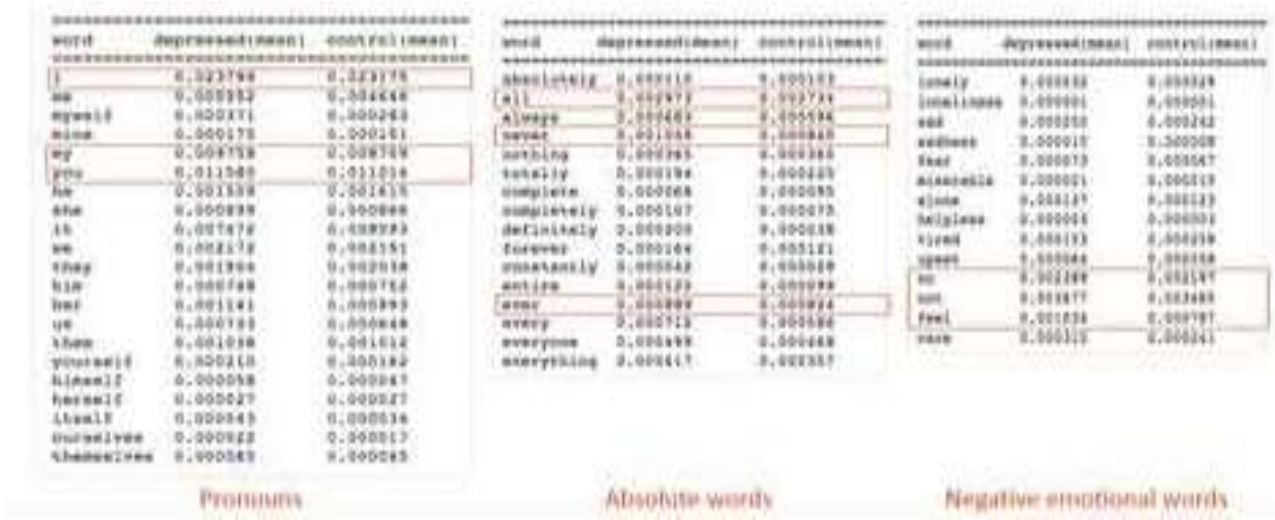

Figure 1: Frequency of words in depressed vs. non-depressed users

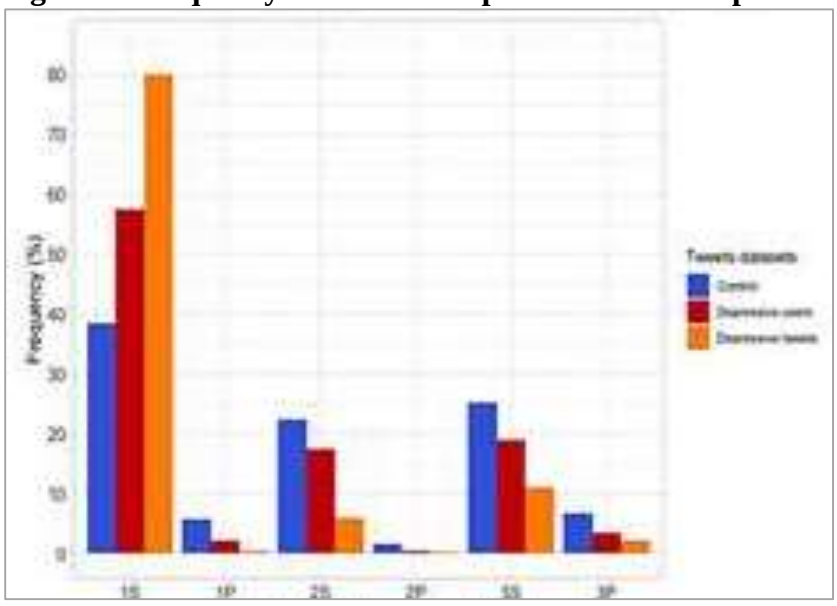

Figure 2: Types of pronouns in depressed vs. non-depressed users. (S = Singular, P = Plural) and (1 = Speaker, 2 = Listener, 3 = Absent)

SentiWord Polarity in Depressed vs. Non-Depressed Users

Using the SentiWord polarity lexicon provided by Python library NLTK, the polarity value (i.e. positive, moderately positive, negative, and moderate negative) of each word was calculated to analyze the emotion of text used by depressed users as compared to non-depressed. From Figure (3) below, it is noticed that mostly depressed users used the percentage of negative and moderate negative words.

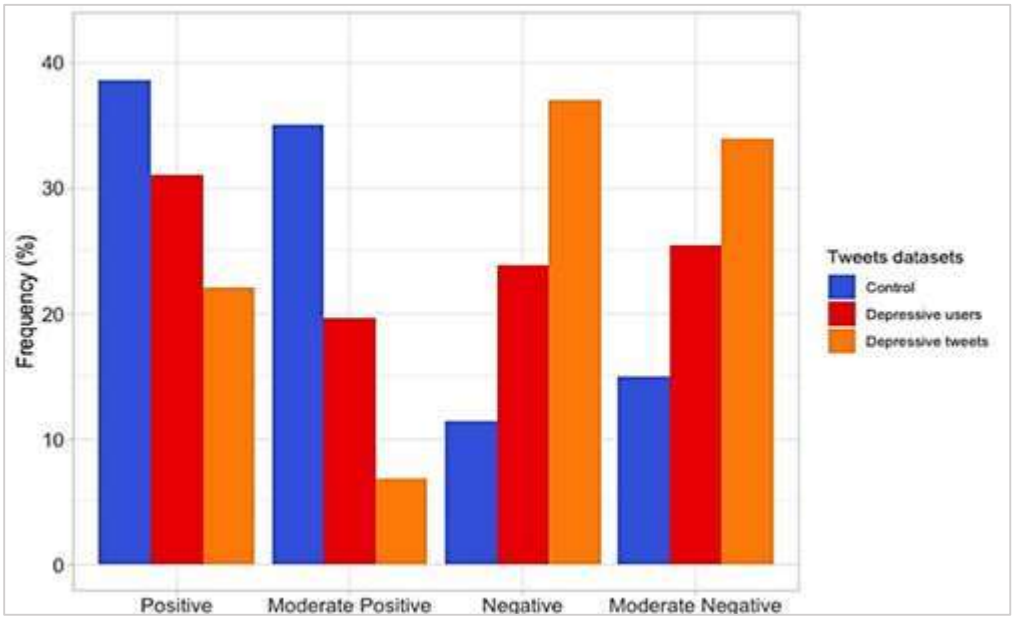

Figure 3: SentiWord polarity in depressed vs. non-depressed users

Word Cloud of Depressed Text (Count Vectorizer)

The third linguistic analysis of the depressive text is the WordCloud of text tokens. This representation will assist in identifying the most used words and phrases by depressed users appearing in large font sizes. This analysis is performed in unigrams and bigrams of words' frequency value and unigrams and bigrams of words' TF-IDF value. From Figure (4) below, it can be noticed that the unigrams with the maximum count value in the depressive text are "feel" and "want". These words do not reflect a specific need or feeling and can be common between depressed and non-depressed users. On the other hand, the bigrams with the maximum count value are "want to die", "good friend" and "feel bad". Now the specific intent of the "feel" and "want" is reflected using bigrams instead of unigrams. However, the bigram "want kill" is a dangerous expression of depression but it is given less importance when using the count value while it is given more importance when TF-IDF value is used, as illustrated in Figure (5) below.

Figure 6: Word cloud of depressing text (count vectorizer)

Figure 7: Word cloud of depressing text (TFIDF vectorizer).

RESULTS

Results of Experiment 1

This experiment evaluates the performance of best scoring baselines selected from the literature on the Twitter dataset of this study. The results of these baselines will be compared against the results of later integration and ensemble experiments.

K+DENCLUE

The application of the K+DENCLUE clustering method proposed in [16] is intended for sentiment analysis. However, when tested on Twitter depression analysis, it has shown less accurate results as illustrated in Table 4 below. The best results on the depression analysis

dataset using this method were achieved when the feature vectors were extracted for chunks of last year tweets compared to individual tweets as applied in the baseline approach. This is because more text holds more discriminative information and context compared to short text. It was observed that the performance of K+DENCLUE proved to be more efficient in terms of cluster count and runtime compared to DENCLUE only. This is due to the advantage of the K-means algorithm, which does clustering given a pre-defined number of clusters followed by DENCLUE clustering for the pre-clustered feature vectors rather than the raw feature vectors.

Table 1: Results of baseline 1 application on depression analysis dataset

Problem Domain	Features Extracted	DENCLUE			K-DENCLUE		
		Accuracy	Clusters	Time	Accuracy	Clusters	Time
Sentiment Analysis	Per Tweet	0.651	17	11 mins	0.629	3	11 mins
Depression Analysis	Last Year Tweets	0.637	23	47 mins	0.621	2	13 mins
	Per Tweet	0.521	61	73 mins	0.574	2	36 mins

SOM+Adam

Similar to the previous approach, the application of Adam deep learning with the SOM clustering method proposed in [17] is intended for sentiment analysis. However, when tested on Twitter depression analysis, it has shown less accurate results as illustrated in Table 5 below. The best results on depression analysis dataset using this method were achieved when bigram TF-IDF features were extracted for chunks of last year tweets per user compared to unigram TF-IDF features extracted for individual tweets as applied in the baseline approach. The results are even higher than the previous K-DENCLUE. The reason is probably that bigrams capture negations and personal pronouns that appear most in depressive text. However, its runtime was much longer.

Table 2: Results of baseline 2 application on depression analysis dataset

Problem Domain	Features Extracted	SOM + PCA + ADL		
		Accuracy	Clusters	Time
Sentiment Analysis	Per Tweet	0.868	3	26 mins
Depression Analysis	Last Year Tweets (unigram)	0.688	2	19 mins
	Per Tweet (unigram)	0.651	2	23 mins
Depression Analysis	Last Year Tweets (bigram)	0.735	2	47 mins
	Per Tweet (bigram)	0.693	2	54 mins

AHDC

The application of the AHDC hierarchical clustering method proposed in [18] achieved 90% accuracy when applied for unsupervised document clustering. However, it has failed when applied for depression analysis dataset, and features were extracted per tweet. For larger document size, i.e. chunks of last year tweets per user, the highest score was achieved which is even higher than K+DENCLUE and SOM+Adam approaches. However, an

accuracy of 73.6% is considered moderate and does not compete with the results supervised approaches reviewed in the literature.

Table 3: Results of baseline 3 application on depression analysis dataset

Problem Domain	Features Extracted	TSG + AHDC
		F-measure
Document Clustering	Per Article	0.9
Depression Analysis	Last Year Tweets	0.736

Clustering approaches, in general, tend to group feature vectors based on un-known and unsupervised similarity criteria. Therefore, when fine-grained similarity criteria are needed, such as depressive of text, clustering algorithms need to be guided to cluster based on that criteria probably by feeding in feature vectors that discriminate and clearly emphasize the difference between depressive and non-depressive text. The following integration of features and ensemble experiments will attempt to achieve enhancement on the performance of the baselines.

Results of Experiment 2

This experiment investigated the integration of multiple combinations of features and document representation methods with the previous baselines clustering approaches. As observed in Tables 7 and 8 and Figures (6) and (7) below, the winner performance of previous baselines is the AHDC clustering when integrated with features of social network (SNF), user profile (UPF), part-of-speech (POS), topic (TPC) and emotion (EMF) along with word2vec embeddings. This combination has achieved an accuracy of more than 79% compared to 73.6% when graph representation and co-occurrence features of words were originally used.

By comparing the accuracy achieved when only topic features were used with the accuracy achieved when other combinations were used, it is noticeable that topic features were the most effective. A similar observation holds with word2vec embedding compared to other document representations. This is due to dimension reduction and high semantic similarity emphasized using of both topic features and word embeddings. Another observation to mentions is that the performance results of unsupervised AHDC clustering were too close from the semi-supervised Adam deep learning with SOM clustering regardless of the feature combination used. The supervised deep learning adds guidance layer to the clustering approach so that the clusters will be generated accurately based on the required criteria. This job is similar to selecting discriminate features to be used with much-unsupervised clustering. The large text was classified more accurately than short text since it holds more information and context.

Table 4: Accuracy results of experiment 2 on Twitter depression analysis dataset

Dataset	Algorithm	Feature	Representation	Accuracy
Twitter	AHDC	SNF + UPF + TBC + POS + EMF	pre-trained Word2Vec	0.79
Twitter	AHDC	TBC + POS + EMF	pre-trained Word2Vec	0.788
Twitter	SOM+Adam	SNF + UPF + TBC + POS + EMF	pre-trained Word2Vec	0.788
Twitter	SOM+Adam	TBC + POS + EMF	pre-trained Word2Vec	0.786
Twitter	AHDC	TBC + POS	pre-trained Word2Vec	0.782
Twitter	AHDC	SNF + UPF + TBC + POS	pre-trained Word2Vec	0.782
Twitter	SOM+Adam	SNF + UPF + TBC + POS	pre-trained Word2Vec	0.78
Twitter	SOM+Adam	TBC + POS	pre-trained Word2Vec	0.78
Twitter	AHDC	TBC	pre-trained Word2Vec	0.78
Twitter	SOM+Adam	TBC	pre-trained Word2Vec	0.778

Table 5: Accuracy results of experiment 2 on Reddit depression analysis dataset

Dataset	Algorithm	Feature	Representation	Accuracy
Reddit	AHDC	SNF + UPF + TBC + POS + EMF	pre-trained Word2Vec	0.7979
Reddit	AHDC	TBC + POS + EMF	pre-trained Word2Vec	0.7959
Reddit	SOM+Adam	SNF + UPF + TBC + POS + EMF	pre-trained Word2Vec	0.7959
Reddit	SOM+Adam	TBC + POS + EMF	pre-trained Word2Vec	0.7939
Reddit	AHDC	SNF + UPF + TBC + POS	pre-trained Word2Vec	0.7898
Reddit	AHDC	TBC + POS	pre-trained Word2Vec	0.7898
Reddit	AHDC	TBC	pre-trained Word2Vec	0.7878
Reddit	SOM+Adam	SNF + UPF + TBC + POS	pre-trained Word2Vec	0.7878
Reddit	SOM+Adam	TBC + POS	pre-trained Word2Vec	0.7878
Reddit	SOM+Adam	TBC	pre-trained Word2Vec	0.7858

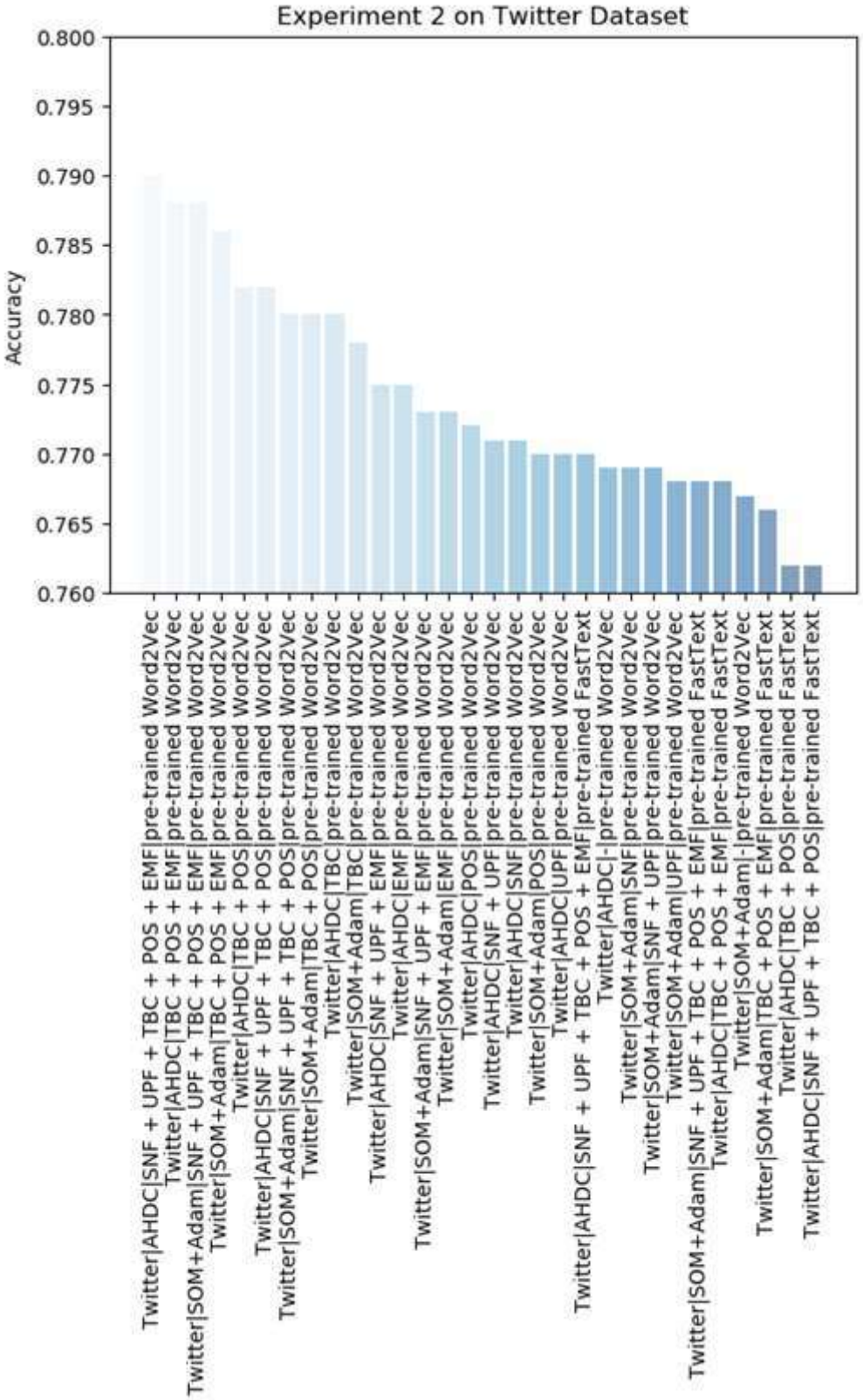
Experiment 2 on Twitter Dataset
Accuracy
0.800
0.795
0.790
0.785
0.780
0.775
0.770
0.765
0.760
Twitter|AHDC|SNF + UPF + TBC + POS + EMF|pre-trained Word2Vec
Twitter|AHDC|TBC + POS + EMF|pre-trained Word2Vec
Twitter|SOM+Adam|SNF + UPF + TBC + POS + EMF|pre-trained Word2Vec
Twitter|SOM+Adam|TBC + POS + EMF|pre-trained Word2Vec
Twitter|AHDC|TBC + POS|pre-trained Word2Vec
Twitter|AHDC|SNF + UPF + TBC + POS|pre-trained Word2Vec
Twitter|SOM+Adam|SNF + UPF + TBC + POS|pre-trained Word2Vec
Twitter|SOM+Adam|TBC + POS|pre-trained Word2Vec
Twitter|AHDC|TBC|pre-trained Word2Vec
Twitter|SOM+Adam|TBC|pre-trained Word2Vec
Twitter|AHDC|SNF + UPF + EMF|pre-trained Word2Vec
Twitter|AHDC|EMF|pre-trained Word2Vec
Twitter|SOM+Adam|SNF + UPF + EMF|pre-trained Word2Vec
Twitter|SOM+Adam|EMF|pre-trained Word2Vec
Twitter|AHDC|POS|pre-trained Word2Vec
Twitter|AHDC|SNF + UPF|pre-trained Word2Vec
Twitter|AHDC|SNF|pre-trained Word2Vec
Twitter|SOM+Adam|POS|pre-trained Word2Vec
Twitter|AHDC|UPF|pre-trained Word2Vec
Twitter|AHDC|SNF + UPF + TBC + POS + EMF|pre-trained FastText
Twitter|AHDC|-|pre-trained Word2Vec
Twitter|SOM+Adam|SNF|pre-trained Word2Vec
Twitter|SOM+Adam|SNF + UPF|pre-trained Word2Vec
Twitter|SOM+Adam|UPF|pre-trained Word2Vec
Twitter|SOM+Adam|SNF + UPF + TBC + POS + EMF|pre-trained FastText
Twitter|AHDC|TBC + POS + EMF|pre-trained FastText
Twitter|SOM+Adam|-|pre-trained Word2Vec
Twitter|SOM+Adam|TBC + POS + EMF|pre-trained FastText
Twitter|AHDC|TBC + POS|pre-trained FastText
Twitter|AHDC|SNF + UPF + TBC + POS|pre-trained FastText

Figure 6: Accuracy results of experiment 2 on Twitter depression analysis dataset

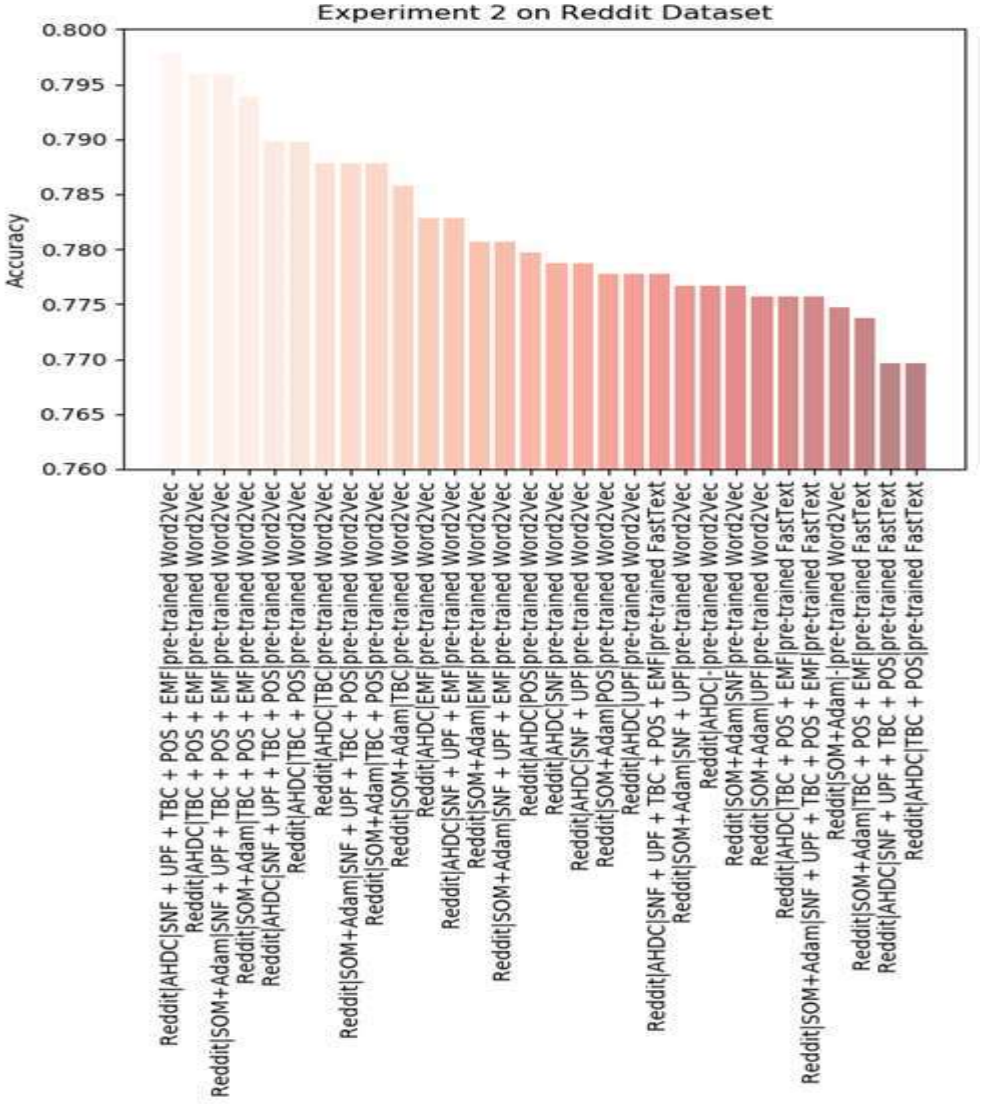

Figure 7: Accuracy results of experiment 2 on Reddit depression analysis dataset

Results of Experiment 3

This experiment investigated the consensus clustering, a.k.a. ensemble learning, when topic features and word2vec embeddings are used. Voting and Stacking ensemble methods were tested. In the voting ensemble, the average prediction results are calculated when two baselines were used in the ensemble, the maximum voting is calculated when three baselines were used otherwise. In stacking, the next baseline will take as input both the original feature vector plus the prediction of the previous baseline is added in a hope to boost the prediction accuracy of the current baseline and hence improve the overall prediction.

By looking at Tables 9 and 10 and Figures (8) and (9) below, it is noticeable that stacking ensembles outperform the voting ensemble. The maximum accuracy achieved is 92.31% when the baselines executed in ascending order of baselines according to performance. That is, when the lowest performance K+DENCLUE was the first, then the second are SOM+Adam then the highest performance AHDC is at the last. This order realizes the fact that the prediction of the weak baseline is boosted by the next stronger baseline. Otherwise, the next weak classifier will degrade the prediction by the strong classifier. However, running time of the top one three-baseline ensemble {K-DENCLUDE -> SOM-Adam -> AHDC} took around 11 hours to finish which is inefficient. To overcome that, one can adopt instead the two-baseline ensemble {SOM+Adam -> AHDC}, which achieved a close accuracy and faster runtime.

Table 6: Accuracy results of experiment 3 on Twitter depression analysis dataset

Dataset	Ensemble	Algorithms	Accuracy
Twitter	Stacking	K-DENCLUDE -> SOM-Adam -> AHDC	0.9231
Twitter	Stacking	K-DENCLUDE -> AHDC -> SOM-Adam	0.9201
Twitter	Stacking	SOM-Adam -> AHDC	0.9147
Twitter	Stacking	AHDC -> SOM-Adam	0.9107
Twitter	Stacking	SOM-Adam -> K-DENCLUDE -> AHDC	0.9017
Twitter	Stacking	AHDC -> K-DENCLUDE -> SOM-Adam	0.8987
Twitter	Stacking	SOM-Adam -> AHDC -> K-DENCLUDE	0.8859
Twitter	Stacking	AHDC -> SOM-Adam -> K-DENCLUDE	0.8847
Twitter	Stacking	K-DENCLUDE -> AHDC	0.8709
Twitter	Maximum Voting	K-DENCLUDE + SOM-Adam + AHDC	0.8709
Twitter	Stacking	K-DENCLUDE -> SOM-Adam	0.8689
Twitter	Average Voting	AHDC + SOM-Adam	0.8679
Twitter	Stacking	AHDC -> K-DENCLUDE	0.8609
Twitter	Stacking	SOM-Adam -> K-DENCLUDE	0.8589
Twitter	Average Voting	K-DENCLUDE + AHDC	0.8129
Twitter	Average Voting	K-DENCLUDE + SOM-Adam	0.8118

Table 7: Accuracy results of experiment 3 on Reddit depression analysis dataset

Dataset	Ensemble	Algorithms	Accuracy
Reddit	Stacking	K-DENCLUDE -> SOM-Adam -> AHDC	0.9323
Reddit	Stacking	K-DENCLUDE -> AHDC -> SOM-Adam	0.9293
Reddit	Stacking	SOM-Adam -> AHDC	0.9238
Reddit	Stacking	AHDC -> SOM-Adam	0.9198
Reddit	Stacking	SOM-Adam -> K-DENCLUDE -> AHDC	0.9107
Reddit	Stacking	AHDC -> K-DENCLUDE -> SOM-Adam	0.9077
Reddit	Stacking	SOM-Adam -> AHDC -> K-DENCLUDE	0.8947
Reddit	Stacking	AHDC -> SOM-Adam -> K-DENCLUDE	0.8935
Reddit	Maximum Voting	K-DENCLUDE + SOM-Adam + AHDC	0.8905
Reddit	Average Voting	AHDC + SOM-Adam	0.8863
Reddit	Stacking	K-DENCLUDE -> AHDC	0.8796
Reddit	Stacking	K-DENCLUDE -> SOM-Adam	0.8776
Reddit	Stacking	AHDC -> K-DENCLUDE	0.8695
Reddit	Stacking	SOM-Adam -> K-DENCLUDE	0.8675
Reddit	Average Voting	K-DENCLUDE + AHDC	0.8377
Reddit	Average Voting	K-DENCLUDE + SOM-Adam	0.8345

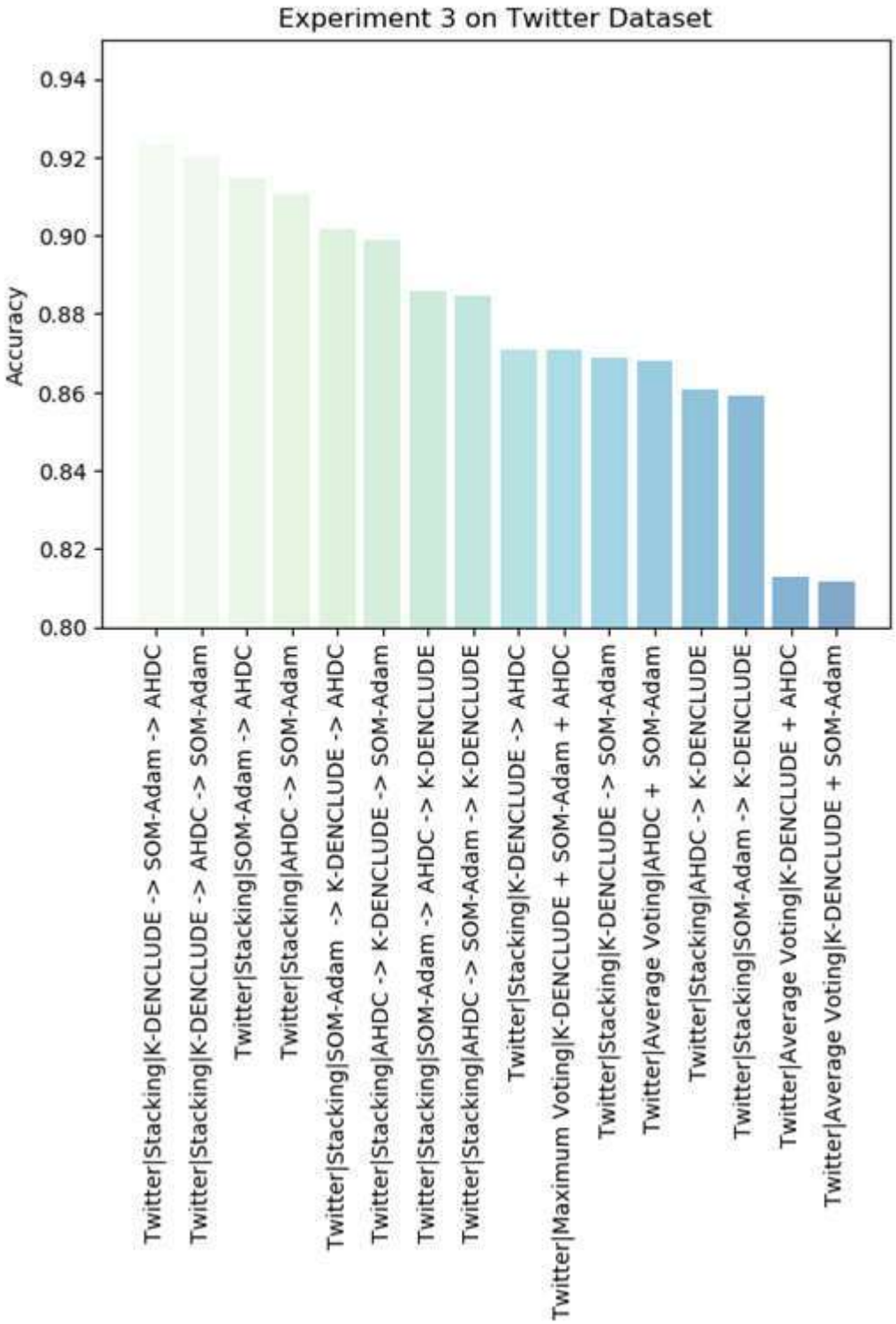

Figure 8: Accuracy results of experiment 3 on Twitter depression analysis dataset

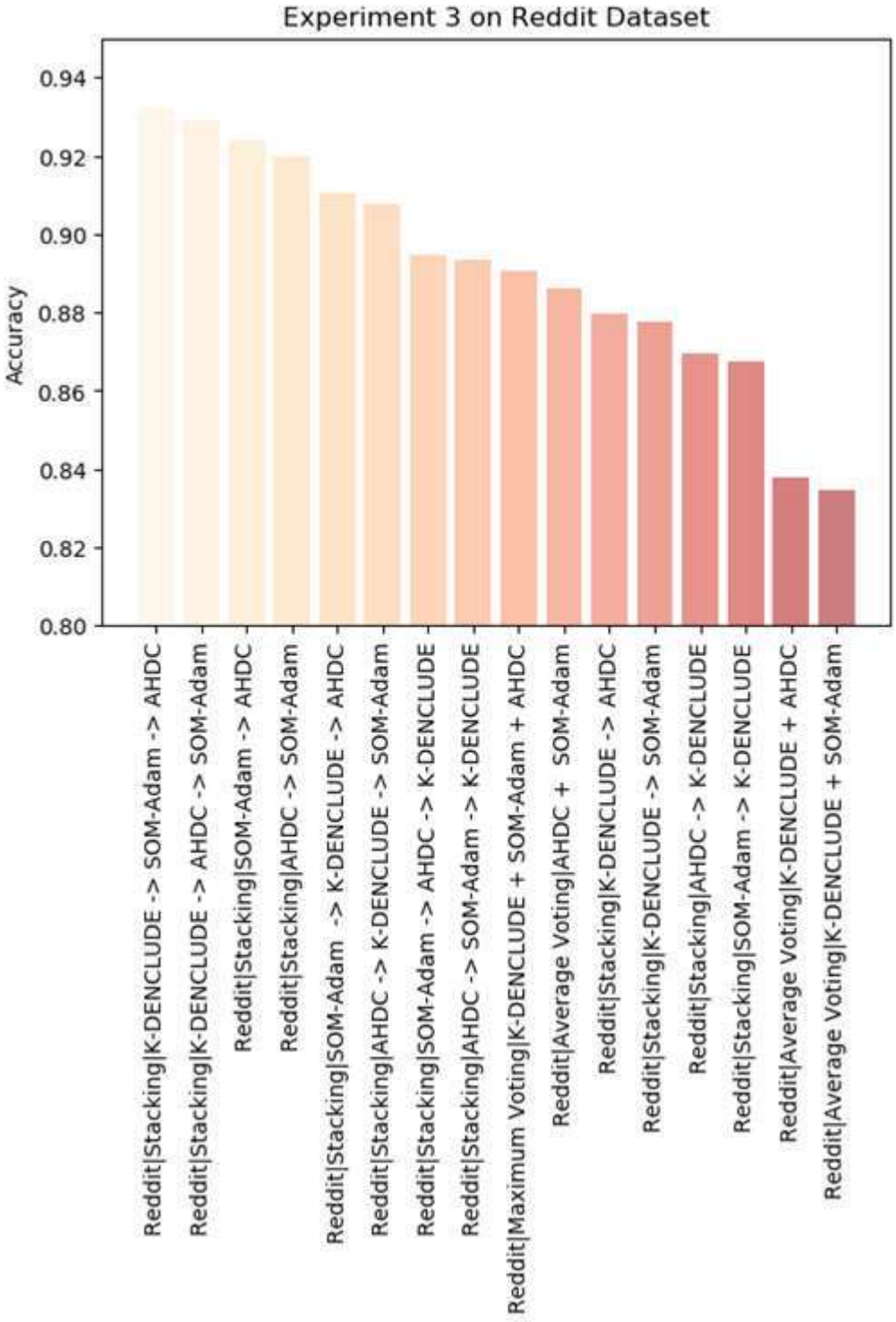

Figure 9: Accuracy results of experiment 3 on Reddit depression analysis dataset

CONCLUSIONS

Feature selection is most important for specific-purpose unsupervised learning. The use of unsupervised and supervised ensembles produces the best scores (like SOM+Adam). The use of topic features and multi-words (semantics) is better than the actual term unigrams. Clinical depression symptoms when used as features result in more accurate results in unsupervised learning. The use of semi-supervised topic modeling such as [13] is highly effective in terms of the quality of learned topics (depressive symptoms). Macroblogs (Reddit) scored higher than microblogs (Twitter). The pre-trained Word2Vec Word embedding scored higher than pre-trained FastText. AHDC clustering scored higher than the rest of the baselines. Stacking Cluster features with the Classification score is close to AHDC clustering. Topic, Emotional, and POS features scored well with clustering.

References

[1] World Health Organization. "Depression", Who.int, 2017. [Online]. Available: https://www.who.int/news-room/detail/30-03-2017--depression-let-s-talk-says-who-as-depression-tops-list-of-causes-of-ill-health

[2] J. Pennebaker, M. Mehl and K. Niederhoffer, "Psychological Aspects of Natural Language Use: Our Words, Our Selves", Annual Review of Psychology, vol. 54, no. 1, pp. 547-577, 2003 [Journal]

[3] J. Pennebaker, R. Boyd, K. Jordan, and K. Blackburn, The development and psychometric properties of LIWC2015. Austin, TX: University of Texas at Austin, 2015. [Software]

[4] J. Hussain et al., "Exploring the dominant features of social media for depression detection", Journal of Information Science, 2019 [Journal]

[5] T. P.N, M. Steinbach and V. Kumar, "Cluster Analysis: Basic Concepts and Algorithms", in Introduction to Data Mining, 2005, pp. 487-568 [Book Chapter]

[6] N. Ismail, "Chapter 9. Social Media and Psychological Disorder", in Social Web and Health Research, 2019, pp.171-192. [Book Chapter]

[7] A. Shatte, D. Hutchinson and S. Teague, "Machine learning in mental health: a scoping review of methods and applications", Psychological Medicine, vol. 49, no. 09, pp. 1426-1448, 2019 [Journal]

[8] B. Norambuena and L. Gutiérrez, "A Systematic Literature Review on Word Embeddings: Proceedings of the 7th International Conference on Software Process Improvement (CIMPS 2018)", in Trends and Applications in Software Engineering, 2019, pp. 132-141.

[9] K. Kalyan and S. Sangeetha, "SECNLP: A survey of embeddings in clinical natural language processing", Journal of Biomedical Informatics, vol. 101, p. 103323, 2020. Available: 10.1016/j.jbi.2019.103323.

[10] A. Orabi, M. Orabi, P. Buddhitha, and D. Inkpen, "Deep Learning for Depression Detection of Twitter Users", in Conference: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, 2018.

[11] G. Shen, "Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution", in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017.
[12] A. Leis, F. Ronzano, M. Mayer, L. Furlong, and F. Sanz, "Detecting Signs of Depression in Tweets in Spanish: Behavioral and Linguistic Analysis", Journal of Medical Internet Research, vol. 21, no. 6, p. e14199, 2019. Available: 10.2196/14199. (116 H-Index, Q1)
[13] A. Yazdavar, "Semi-Supervised Approach to Monitoring Clinical Depressive Symptoms in Social Media", in Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, Australia, 2017 [Conference Proceedings]
[14] S. Tseng, Y. Lu, G. Chakraborty and L. Chen, "Comparison of Sentiment Analysis of Review Comments by Unsupervised Clustering of Features Using LSA and LDA", 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), 2019 [Journal]
[15] M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of Depression-Related Posts in Reddit Social Media Forum", IEEE Access, vol. 7, pp. 44883-44893, 2019 [Journal]
[16] H. Rehioui and A. Idrissi, "New Clustering Algorithms for Twitter Sentiment Analysis", IEEE Systems Journal, pp. 1-8, 2019 [Journal]
[17] M. Ali, M. Sarowar, M. Rahman, J. Chaki, N. Dey, and J. Tavares, "Adam Deep Learning With SOM for Human Sentiment Classification", International Journal of Ambient Computing and Intelligence, vol. 10, no. 3, pp. 92-116, 2019 [Journal]
[18] R. Nagarajan, "Document Clustering Using Agglomerative Hierarchical Clustering Approach (AHDC) And Proposed TSG Keyword Extraction Method", International Journal of Research in Engineering and Technology, vol. 05, no. 11, pp. 118-124, 2016 [Journal]
[19] Y. Wang, K. Kim, B. Lee, and H. Youn, "Word clustering based on POS feature for efficient twitter sentiment analysis", Human-centric Computing and Information Sciences, vol. 8, no. 1, 2018 [Journal]
[20] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space", in International Conference on Learning Representations, 2013 [Conference Proceedings]
[21] M. Nadeem, "Identifying Depression on Twitter", 2016. [Archive Material].