SentiNeural: A Depression Clustering Technique for Egyptian Women Sentiments

Doaa Mohey ElDin[1]
Information System Department
Faculty of computers and Information, Cairo University
Cairo, Egypt

Mohamed Hamed N. Taha[2], Nour Eldeen M. Khalifa[3]
Information Technology Department
Faculty of computers and Information, Cairo University
Cairo, Egypt

Abstract—Online Sentiments Analysis is a trending research domain of study which is based on natural language processing, artificial intelligence, and computational linguistics. Negation sentiments usually are not included in sentiment's analysis process. The depression analysis can be improved by negative sentiments processing. The negation sentiments may contribute to classify the depression problems and its causes. The proposed clustering technique can detect female sentiments from the sentiment's text through cause's classification, and the written sentiment style. The combination of sentiment analysis and neural network is a promising solution for creating a new clustering algorithm. According to Egypt Independent Journal in 2018, 7% of Egyptians suffer from mental illness reported by the Public Health Ministry in Egypt. But the real statistics is more than the mentioned percentage which causes major social problems such as divorce, avoiding responsibilities, or non-marriage. This paper will address the real statistics and cluster the depression causes and social status for each sentiment. Online women sentiments are the essential focus of this research. The proposed technique consists of two algorithms clustering for user's sex and classification algorithm for causes and responsibilities of women. The proposed clustering algorithm can recognize automatically for the sentiments user sex (females or males) and the level of depression automatically. The neural network clustering approach will produce accurate analysis results. The hardness of depression analysis implicitly and explicitly demonstrated in the different classifications for sentiments. This paper introduces a new technique for clustering sentiments and evaluating Egyptian women depression based on social sentiments.

Keywords—Sentiment analysis; negation handling; depression analysis; neural network; clustering

I. INTRODUCTION

Recently, women spend a lot time online on social networks and communities. That becomes a main platform for expressing their feelings such positive or negative. The depression is one of the hardest text classification and recognition analysis. This research focuses on clustering the sentiment's texts from women or men. This paper takes care of the women sentiments and opinions, so that this technique can cluster the women' sentiments based on the written style, automatic responsibilities identification, and causes classification. The classification of responsibilities consists of a disease, and financial status. According to Egypt Independent Journal in 2018, 7% of Egyptians suffer from mental illness reported by the Health Ministry in Egypt[1].

More solid and real statistics can be obtained by analyzing the women sentiments on social media networks.

Most the depression analysis is based on age or work, but this research based on responsibilities and work for women in Egypt. After the horrific reports announced by the World Health Organization [2], According to Global Terrorism Database, the number of suicide bombers exceeded 400 attacks per year in Egypt [3], in addition to thousands of suicide attempts in the homes and streets of Egypt every year. About half of these attempts and suicides are from ladies or women.

The share of women in the labor force, or as some call the labor market, accounts for almost a third of women's participation: 30.8% are female, 21.2% are male, and 118.9 thousand women in the age group (12-17 years old) have married (1.2 thousand widows, 1.2 thousand divorcees, 111 thousand married women, 5.5 thousand marriages) and will show population estimates for 2017. The rate of unemployment between normal and 23.6% compared to 8.9% among males in 2016, and the proportion of those who work in permanent work 76.5% of the total working factor compared to 71.1% of males among the total working male. 17.6% compared to 4.5% for health insurance versus 64.1% compared to 45.4% for males [4].

Women responsible for children and spend money on them by 16%. The highest divorce rate for divorcees was less than 25 years in (2010) by (22.58%). Percentage of divorce to total divorce certificates in 2010 (3335 cases) 2.2%. More than 77% are workers at home and out home. These responsibilities can keep women in depression although denying more than 20% about depression. But they suffer deep problems in their work, home, and relationships [4]. So, this paper presents a classification algorithm for identifying the causes, and responsibilities recognition and sentiments evaluation on Facebook social network.

Sentiment Analysis becomes a recent research trend for improving software development, decision making. Social media can be an essential source of sentiments tweets, app reviews, bloggers comments. Sentiment analysis is defined by the subjectivity study (neutral vs. emotionally loaded) and class polarity (positive vs. negative) of a sentiment word [5].

This paper presents a proposed technique for evaluating the sentiment analysis of Egyptian women. Using Artificial Intelligence [6], machine learning[7], and deep learning

algorithms [8], it also can create a new clustering algorithm for written style for women based on written style and causes classification from Facebook reviews. The proposed technique can provide the identification social status and financial status for each woman that can be a parameter in the relationship between depression and causes.

The rest of this paper is organized as follows: Section 2 presents related works. Section 3 shows the proposed framework. In Section 4, experiment and discussion will be illustrated. Finally, Section 5, conclusion and proposes directions for future work will be presented.

II. BACKGROUND

This section presents a summary of the essential aspects of sentiment analysis and discusses the initial research efforts in the sentiment techniques and applications.

A. Sentiment Analysis

It expresses the mining of users' reviews [5 & 9] about product/ topic online as a social network or blogger. The recent research introduces the challenges [10] of sentiment meaning understanding and sentiment parsing sentences and words. Writer's review is a major criterion for the quality of services enhancement to grant deliverables. There are multisource of these reviews as social networks, web portals, and Blogs. Recently, the researches targets to analyze and benefit online sentiments not only on sentiment polarity positive or negative but also on the analyzing the variant feelings as depression, violent, risky, or serious for negative sentiments and wishing, safety, happiness for positive analysis.

B. Sentiment Classification & Clustering

There are difference between classification [10] and clustering [11] for sentiment 'text. Sentiment Classification is a type of learning models for identifying data classes and pattern recognition based on labeled data. It depends on the predefined data or classes. There are several motivation in classification for supporting the results in various data types (as text, images, or videos) and domains (such as medical, tourism, online reviews, etc…). Clustering is considered a learning technique model also but it is based on unknown classes previously. It can support the objects and relationships automatically based on the input data. It is based on similarity between data and objects. It is known unsupervised technique. This clustering targets determining the relations between different sentiments, distributing them dynamically in natural groups, or discovering the most relevant subjects within their content and expressing them in their own terms

C. Deep Learning

It is an important branch of machine learning. Deep learning [8, 12 & 13] has several algorithm to improve accuracy and performance. That can support the relationships and discover the hidden features between layers of data.

D. Sentiment Summarization

The summarization of sentiment aims at generating a summary [14] of entities and opinion analysis that is representative of the average opinion and speaks to its important aspects. A standard sentiment setting for social networks summarization assumes a set of sentiments S = {s1, . . . , sn} that include reviews about interesting topic or something. The target of any summarization system is to generate a summary S of that topic that is representative of the average review and talks to its significant aspects. The summarization requires to measure with respect to the mentioned aspects: optimization, and intensity.

The optimization is measured by the equation (1):

$$argmax\ L(SM)\ , \text{s.t.}: Length(SM) \leq K \tag{1}$$

Where L is many possible scores of summaries, Length (SM) refers to all sentiments summary length and K is the pre-specified length constraint.

The intensity of summarization is declared in equation (2):

$$Intensity\ (SM) = \sum_{t \in SM} | LEX - S\ (t)| \tag{2}$$

$$S\ (SM) = \frac{\sum_{t \in SM} Lex - S\ (t)}{\alpha + intensity\ (SM)} \tag{3}$$

A central function for normalized sentiment,

III. PROPOSED CLUSTERING TECHNIQUE

This study focuses on four kinds of factors such as sentimental process, personal style process, linguistic style, and negative sentiment depression level. The proposed technique consists of two algorithms clustering for user's sex and classification algorithm for causes and responsibilities of women. There are some features extracted from this classification that can provide results by identifying the social status and financial status for each woman that can be a parameter in the relationship between depression and causes. The proposed classification algorithm is based on six classes (education problems, work problems, family problems, baby problems, disease problems, and shape appearance/personal problems). The social status (single, married, divorced, married and has children, widowed). This algorithm shows the evaluation and classification for each sentiment based on deep neural algorithms (word2vector embedding algorithm and recurrent neural network). Word embedding that is introduced by word2vec is generally used to learn context and produce high-dimensional vectors in a space. These embedding are then classified using the machine-learning algorithm.

These features can be analyzed and detect the depressive information for each sentiment that enable to recognize some new clusters that can detect the writer' sex based on the sentiment written style and some causes classification which extracted from online sentiments data such as (causes of depression, responsibilities from each cause, social and finical status). These extracted sentiments received as Facebook posts or comments. We then apply supervised deep learning approaches to considerate each factor kinds independently. The used classification technique is Word2vec and LSTM. The clustering algorithm can produce the man, and female sentiments (see Fig. 1). This research focuses on women depression so that is not easy to cluster the sentiment based on the written requirements and written style.
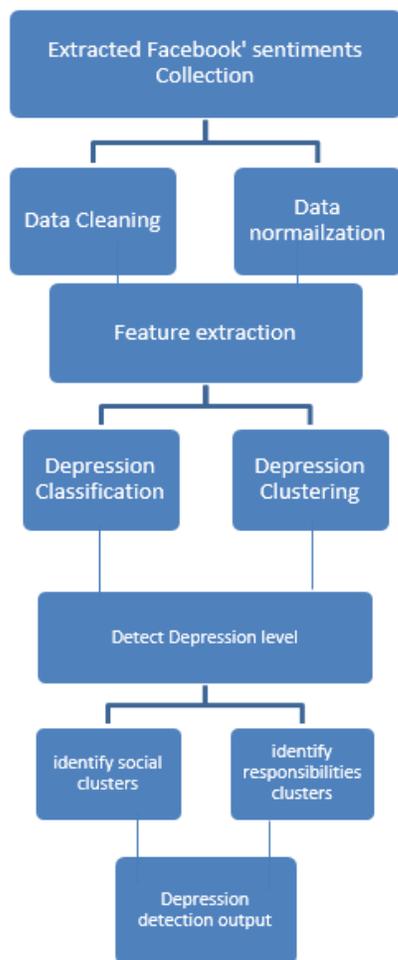
Fig. 1. The Proposed Technique of SentiNeural (Holding a Proposed Clustering and Classification Algorithms).

Feature extraction support describing and demonstrating amongst depressive and non-depressive posts. Fetching many features includes linguistic specifications from user's reviews. It is clarified briefly as follows:

- Psychological analysis operation —affective process, social process, cognitive process, time orientations, relativity, personal concerns

- Part-of-speech Linguistic operation such as number of words on sentiments sentence, word, pronoun, prepositions, adverbs, conjunctions, Negations.

- Other types of language grammar such as verbs, adjectives, adverbs, comparisons, numbers, or quantifiers.

This research applies text analysis algorithms in sequenced processes as follow:

A. Translation

This algorithm is suitable to variant language but this paper focuses on the Arabic language dataset. Translate Arabic into English process: the proposed algorithm works for English sentiments, so we will translate the Arabic language into English and works on the English sentiment. The translation has the challenge to detect language and slang words. The slang words interpretation in Arabic Language requires several motivations for reaching better results. That is considered a new research opportunity.

B. Preprocessing Phase Remove stop Words

A stop word is a commonly used word (such as "the", "a", "an", "in") that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. Part-of-speech tagging (POST) is considered grammatical tagging for sentiment text. Stemming: A stemming algorithm is a process of linguistic normalization, in which the variant forms of a word.

C. Hierarchal Classification Sentiments

This clustering is a new proposed approach for sentiment analysis. It creates a hierarchal clustering model for detecting women sentiments automatically based on the sentiment style of written and a new depression classification for causes and responsibilities. For each sentiment S that requires to identify the polarity first such as (positive or negative sentiments). The technique avoids the positive sentiments and focuses on negative sentiments. The proposed technique splits negative sentiments into three levels of negative sentiment (high, medium, and low levels of depressions). Clustering [11, 12] can be roughly featured as Hard clustering: each object belongs to an exact cluster or not. Soft clustering (also: fuzzy clustering): each object belongs to each cluster to a certain degree (for example, a likelihood of belonging to the cluster).

It also can predict causes using machine learning and deep learning algorithms (word2vec embedding algorithm and Recurrent neural network). Word2vec is known aslo Word embeddings are basically a form of word representation that bridges the human understanding of language to that of a machine. Word embeddings are distributed representations of text in an n-dimensional space. These are essential for solving most NLP problems. The main goal for W*ord2vec* algorithm is applying on big data of text and generating a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space.

This cluster algorithm can recognize automatically for the sentiments user sex (females or males) and the level of depression. This research takes care of women/ female sentiments on Facebook social network which becomes one of the *biggest platform for expressing feelings and opinions simultaneously.

D. A Depression Clustering

This cluster algorithm can recognize automatically for the sentiments user sex (females or males) and the level of depression. It is based on LSTM neural network as the following in Fig. 2.
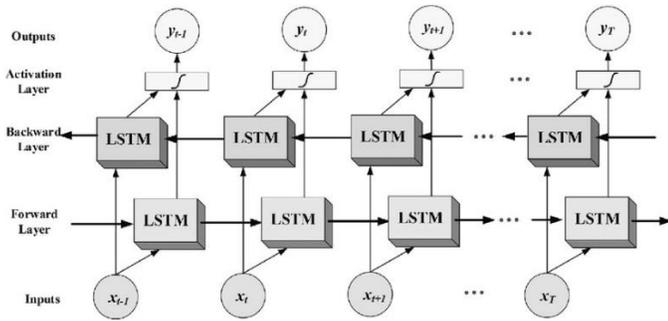
Fig. 2.   The Neural Network Structure for the Depression Analysis.

E. Sentiment Evaluation

That can detect the sentiment polarity to can determine the sentiment is positive or negative. We focus on this research on the negative polarity sentiments. The level of negative refers to the level of depression. Sentiment evaluation: this process requires the text analysis, interpretation, evaluate the words, sentence, and sequence of words. Sentiment polarity: is based on positive or negative polarity. Sentiment depression level: is based on negative sentiment polarity.
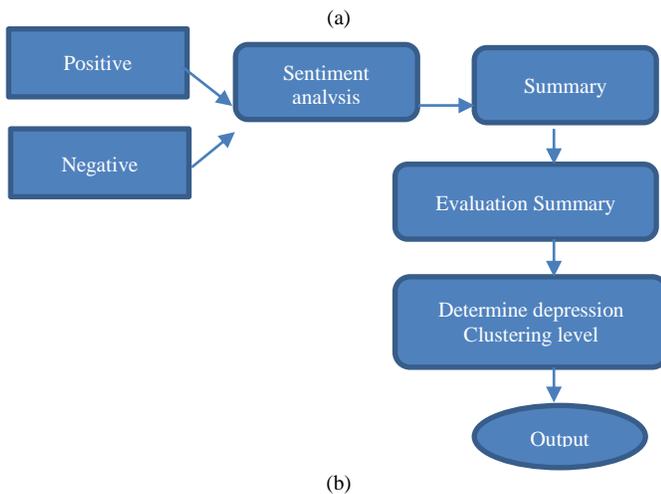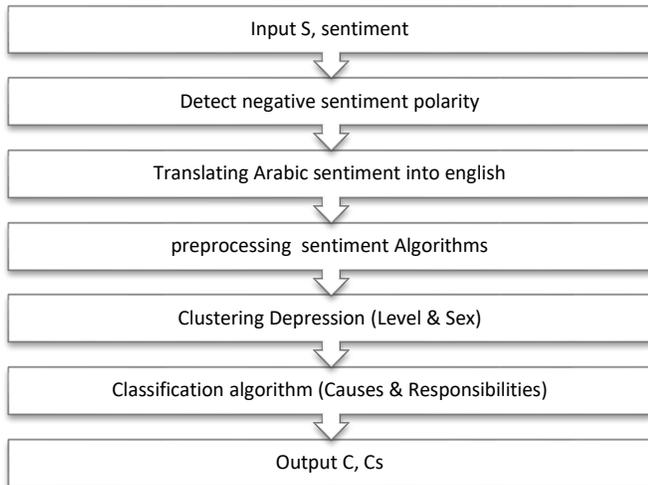

(a)

(b)

Fig. 3.   The Proposed Depression Detection Technique (a) The Sequenced Processes of the Technique (b) The Relationship between Sentiment Analysis, Summary and the Results of Depression Level.

TABLE I.   THE PROPOSED CLUSTERING SENTIMENT BASED ON NEURAL ALGORITHM

A Proposed Algorithm of : for Arabic Sentiments	
Input	A corpus S of N number of Sentiments {s1, s2, . . . , sN}
Output	Assign a Positive OR Negative label For each sentiment si ∈ S, (i =1, 2, . . . , m)
Pre-processing:	1: for all sentiments sj ∈ S do 2: for all each token ti ∈ Sj do (token refers to the word) 3: Tag ti with part of speech tagging ptj 4: if ptj == %x OR ptj == x%, x donates searching to the infinitive verb from the token whether noun, verb, adjective and adverb respectively then 5: Keep pti 6: Add pti to F, F is the features set 7: else 8: Remove ti 9: end if 10: end for 11: end for
	Clustering: 12: Set the clusters number C = 2 13: for all matrix files Mi , (i = 1, 2, . . . , n), do 14: using Deep neural network algorithm, learning the features and properties of women feelings problems based on word2vec embedding algorithm. 15. Determine the negative polarity classification for the translated sentiment. 16. if sj ∈ Ns do, Ns refers to Negative sentiment 17: Cluster Lj into three levels of depression based on recurrent neural network algorithm L1,L2,L3
Evaluation	18. Evaluate the Sp into three classifications of negative sentiment polarity from -1, to 0, -1 refers to high depression, -.05 medium depression, and 0 low depression.
A proposed clustering algorithm for Facebook 'users	19. Cluster sj into two user's sex male M or female F 20: Predict sj cluster having depression or not 21: Predict depression level into three clusters (high, medium, low) 22: endif 23: end for
A proposed classification algorithm for reasons of depression	24. Classify sj into five social clusters G1, G2,G3,G4, and G5 25: Predict sj class three financial for each sentiment 26: Predict class the cause problem ci (from six classes) 27: Predict the responsibility class 28: endif 29: end for
Poisson Distribution	30. counting data using poison regression, which refers to the probability of events for a Poisson distribution $P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$ (1) , λ is the mean number of events x=0,1, 2....

F. Sentiment Summarization

It targets a summary of causes of depression based on the level of depression and responsibilities.

G. Poisson Distribution Phase

This distribution targets the determination of the number of depression women from the collected dataset. It aims to discuss the number for occurrences the depression in other conditions (multi-clustering) e.g. high depression, married, and cancer. A simple flowchart for the phases of the proposed framework is presented in Fig. 3, while Table I shows the pseudocode of the proposed framework.

IV. EXPERIMENT AND RESULTS

This experiment applies on Arabic sentiments for women on social media.

A. Preparing the Dataset

The dataset is manually collected from women groups on Facebook. We observe the users' reviews for depressive behavioral inspection and detection. The social network includes the other faces and depth emotions of people that can express about them in a freeway. Preparing of social network data, in particular Facebook writer's reviews is one of the main obstacles which give information on whether or not they could contain depression producing content. That is not easy to use any extractor program for recognizing the Arabic sentiment from women. So the preparing the data is based on the clustering sex male or women online. We depend on Facebook 'groups in Egypt.

After gathering the raw data from Facebook, the analyzing reviews process is applied on the sentiment analysis word level. It can interpret text and prepare it for evaluation. Our primary dataset contains 10.000 sentiments (7000 negatives and 3000 positives) from Facebook from groups to ensure clustering women from without examining their profiles.

Table II shows a sample of women sentiments in the first phase (Translation from Arabic into English), and Table III shows SentiNeural results in processing in the Multi-clustering based on deep neural networks for achieving the causes of depression based on clustering (financial status, social status, and responsibilities).

B. Accuracy of Sentiment Analysis Polarity

The women sentiment will be evaluated accuracy in two times. First time to know the positive or negative sentiments is shown in Table IV. The second time is evaluated for the classification levels of negative as shown in Table V.

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn} \qquad (4)$$

TABLE II. WOMEN SENTIMENTS TRANSLATION FROM ARABIC TO ENGLISH

No	Sentiment	Translate
1.	المدير متحكم وودني وكل شوية يقلل من شغلى مش عارفة اثبت نفسي أزاي	The manager is controlled and Woody and every angle reduces my job not knowing how to prove myself
2.	الشغل في الشركة صعب جدا مش بعرف اخد بريك خالص وبتاخر ل 9 او ل كل يوم103 بالي	The work in this company is very heavy I can't take a break and must to go out around 9 or 10 daily
3.	انا تخينة اوي وكل الناس بتقول عليا وحشة وزني فوق ال130 كيلو مش عارفة اعمل ايه	I am overweight more than 130 kilo, I don't know what can I do
4.	لو سمحتوا ابني تعبان اوي وسخن اعمل ايه مش معايا فلوس اروح لدكتور	If you allow my son is very tired and hot I do not know how to make money
5.	من وقت ما اتطلقت وامي تضربني وتهيني كأني عالة عليهم	From time to time my mother tasted me and humiliated me as if I was a burden on them

TABLE III. MULTI-CLUSTERING FOR PREDICTING THE CAUSES OF DEPRESSION INTO TO LEVELS OF CLUSTERING FROM EXTRACTING FEATURES IN WOMEN SENTIMENTS FOR THE PREVIOUS EXAMPLES IN TABLE I

No.	Social status cluster	Financial status cluster	Causes	Responsibilities
1.	Single	Medium	Manager problem	Work
2.	Single	Medium	Work dates and delays	Work
3.	Single	Medium	The shape and weight	Non
4.	Married and has child	Low	Baby patient	Baby
5.	Divorce	Low	Divorce and mother problems	Divorce

TABLE IV. THE ACCURACY OF POSITIVE OR NEGATIVE POLARITY

	Precision	recall	F1-score
Negative	0.97	0.85	0.91
Positive	0.67	0.80	0.83
Average	0.85	0.82	0.83

TABLE V. THE ACCURACY FOR NEGATIVE LEVELS

	Precision	recall	F1-score
High negative	0.77	0.65	0.71
Medium negative	0.41	0.52	0.46
Low negative	0.47	0.60	0.53
Average	0.65	0.62	0.63

Tp refers to true positive, Fn: false negative, Tn: true negative, and Fn: false negative

C. Clustering Machine Learning and Poisson Distribution

The Poisson regression algorithm [15] is used to predict numbers based on regression models. The response variable has a Poisson distribution. The distribution of regression refers to a discrete distribution that is a method with non-whole numbers. According to Fig. 4 and Fig. 5, the results are found: Mean Absolute Error= 49958.764234, Root Mean Squared Error = 52002.038613, Relative Absolute Error =3.996701, Relative Squared Error= 12.980218, and Coefficient of Determination= -11.980218.

These results discuss the relationship between responsibilities, or causes, financial status/level, and diseases. Note: without caring the age because there are responsibilities can be in a different age, (for example, marriage age, work age, experience age, born children age, poor age, and disease age).
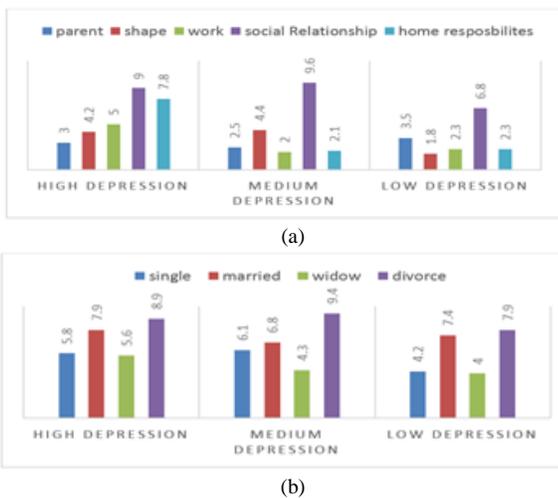
(a)

(b)

Fig. 4. (a) The Poisson Regression Illustrates the Relationship between the Levels of Depression and the Causes.(b) The Poisson Regression Illustrates the Relationship between the Levels of Depressions and the Predicting Social Status.
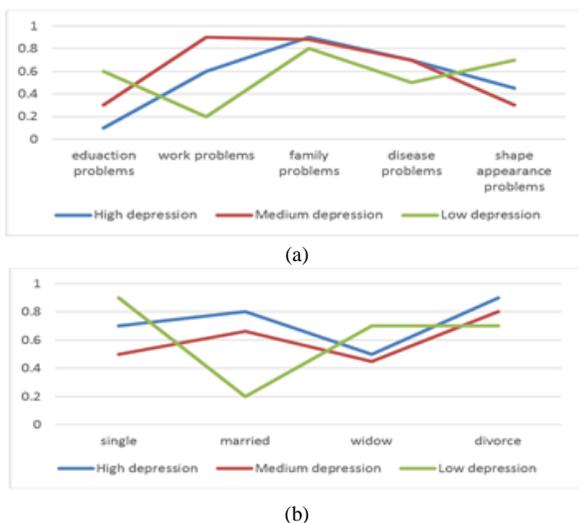

(a)

(b)

Fig. 5. (a) The Poisson Regression Distribution Shows for the Hierarchal Clustering of Translated Sentiments into Causes Clustering Dimension (Education Problems, Work Problems, Family Problems, Baby Problems, Disease Problems, and Shape Appearance Problems). (b) The Poisson Regression Distribution Shows for the Highest Rate of Family Problems Clustering of Translated Sentiments (Single, Married, Widow, and Divorce).

V. CONCLUSION

Online sentiments have a big effect in decision making in business. There are several challenges in analyzing and evaluating sentiments. More than 60% of sentiments face a negative polarity challenge. This paper proposes SentiNeural which is a new clustering and evaluating online women sentiments from Facebook. This technique targets clustering the user's sentiments based on the text of sentiments style and clustering the level of depression based on cause's classification algorithm. SentiNeural introduces a new classification algorithm for extracting the causes, responsibilities, financial status, and social status for each women using deep neural algorithms. It also includes a

translation from Arabic to English languages and shows a summary for each sentiment. The experiment of the proposed framework relies on Poisson regression distribution that can determine the number of sentiments in (high, medium, and low) depression according to the different causes (family, education, work, diseases, baby, shape/personal, problems) with respect to the prediction of two levels finical and social status (married, single, widow, divorce). The accuracy reaches in average between 85% into 91% that is based on translation and summarization results. For future work, improving the translation and summarization algorithms for achieving 98% in accuracy. Further, another future work targets Appling the same algorithm in various languages and providing some solution for Arabic language challenges.

REFERENCES

[1] Egypt Independent, "7% of Egyptians suffer from mental illness: Health Ministry," Egypt Independent, 2018. [Online]. Available: https://egyptindependent.com/7-of-egyptians-suffer-from-mental-illness-health-ministry/. [Accessed: 18-May-2019].

[2] "Depression and Other Common Mental Disorders : Global Health Estimates," Organization, World Health, 2018. [Online]. Available: https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf?sequence=1. [Accessed: 18-May-2018].

[3] N. E. M. Khalifa, M. H. N. Taha, S. H. N. Taha, and A. E. Hassanien, "Statistical Insights and Association Mining for Terrorist Attacks in Egypt," 2019, pp. 291–300.

[4] E. Ortiz-Ospina and S. Tzvetkova, "Working women: Key facts and trends in female labor force participation," Our World in Data, 2017. [Online]. Available: https://ourworldindata.org/female-labor-force-participation-key-facts. [Accessed: 18-May-2019].

[5] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks," arXiv Prepr. arXiv1410.8586, 2014.

[6] W. B. Gevarter, "Introduction to Artificial Intelligence.," Chem. Eng. Prog., 1987.

[7] Y. Baştanlar and M. Özuysal, "Introduction to machine learning," Methods Mol. Biol., 2014.

[8] F. Q. Lauzon, "An introduction to deep learning," in 2012 11th International Conference on Information Science, Signal Processing and their Applications, ISSPA 2012, 2012.

[9] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media.," in Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media 128, 2013, pp. 128–137.

[10] G. Shen et al., "Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution," in Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017, pp. 3838–3844.

[11] Y. Cao, M. Huang, and X. Zhu, "Clustering sentiment phrases in product reviews by constrained co-clustering," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2015.

[12] N. Eldeen, M. Khalifa, M. Hamed, N. Taha, and A. E. Hassanien, "Aquarium Family Fish Species Identification System Using Deep Neural Networks," 2018.

[13] N. E. Khalifa, M. Hamed Taha, A. E. Hassanien, and I. Selim, "Deep galaxy V2: Robust deep convolutional neural networks for galaxy morphology classifications," in 2018 International Conference on Computing Sciences and Engineering, ICCSE 2018 - Proceedings, 2018, pp. 1–6.

[14] P. Beineke, T. Hastie, C. Manning, and S. Vaithyanathan, "Exploring Sentiment Summarization," Proc. AAAI Spring Symp. Explor. Attitude Affect Text Theor. Appl., 2004.

[15] P. Sedgwick, "Poisson regression," BMJ (Online). 2014.