APPLIED COMPUTER SCIENCE

The Journal is a peer-reviewed, international, multidisciplinary journal covering a broad spectrum of topics of computer application in production engineering, technology, management and economy.

The main purpose of Applied Computer Science is to publish the results of cutting-edge research advancing the concepts, theories and implementation of novel solutions in computer technology. Papers presenting original research results related to applications of computer technology in production engineering, management, economy and technology are welcomed.

We welcome original papers written in English. The Journal also publishes technical briefs, discussions of previously published papers, book reviews, and editorials. Especially we welcome papers which deals with the problem of computer applications in such areas as:

- manufacturing,
- engineering,
- technology,
- designing,
- organization,
- management,
- economics,
- innovations,
- competitiveness,
- quality and costs.

The Journal is published quarterly and is indexed in: BazTech, Cabell's Directory, CNKI Scholar (China National Knowledge Infrastucture), ERIH PLUS, Index Copernicus, J-Gate, Google Scholar, TEMA Technik und Management.

Letters to the Editor-in-Chief or Editorial Secretary are highly encouraged.

CONTENTS

Applied Computer Science, vol. 16, no. 1, pp. 5–17
doi:10.23743/acs-2020-01

Submitted: 2019-10-12
Revised: 2019-11-25
Accepted: 2020-03-02

Electricity Demand Forecasting, STLF,
Deep Learning Techniques, LSTM, CNN, MLP

Saheed A. ADEWUYI[*], *Segun AINA* [0000-0001-5080-1760][**],
Adeniran I. OLUWARANTI [0000-0003-4920-6053][**]

A DEEP LEARNING MODEL FOR ELECTRICITY DEMAND FORECASTING BASED ON A TROPICAL DATA

Abstract

Electricity demand forecasting is a term used for prediction of users' consumption on the grid ahead of actual demand. It is very important to all power stakeholders across levels. The power players employ electricity demand forecasting for sundry purposes. Moreover, the government's policy on its market deregulation has greatly amplified its essence. Despite numerous studies on the subject using certain classical approaches, there exists an opportunity for exploration of more sophisticated methods such as the deep learning (DL) techniques. Successful researches about DL applications to computer vision, speech recognition, and acoustic computing problems are motivation. However, such researches are not sufficiently exploited for electricity demand forecasting using DL methods. In this paper, we considered specific DL techniques (LSTM, CNN, and MLP) to short-term load forecasting problems, using tropical institutional data obtained from a Transmission Company. We also test how accurate are predictions across the techniques. Our results relatively revealed models appropriateness for the problem.

[*] Osun State University, Department of Information and Communication Technology, Osogbo, Osun State, Nigeria, saheed.adewuyi@uniosun.edu.ng
[**] Obafemi Awolowo University, Department of Computer Science and Engineering, Ile-Ife, Osun State, Nigeria, s.aina@oauife.edu.ng

1. INTRODUCTION

Electricity demand forecasting is a concept in the power system that is used to describe prediction of the load on the power grid ahead of actual consumption. The power system grid must always bear the demand and thus satisfactorily service its customers. Electricity is an ever demanded commodity due to geometric rise in population (IAE, 2018). In Nigeria, a typical tropical climate, population is a key factor hindering the expected equilibrium of supply and demand of electricity as a commodity, among other factors. Therefore, there exists a gap between the generated power and consequential distribution of the commodity in Nigeria. Although, it was learnt that there is also a problem of carrying capacity of the transmission lines which contributes to supply shortage of electricity to users. As a result, the power stakeholders subject its users to harsh situations of inadequate or epileptic power supply or even blackout sometimes. While keeping this scourge under control, the distribution company will nonetheless need to always carry out demand or load forecasts of electricity consumption ahead of time.

Electricity demand forecasting or load forecasting is categorised into three main groups, generally. This includes Short-Term Load Forecasting (STLF), Medium-Term Load Forecasting (MTLF) and Long-Term Load Forecasting (LTLF) (Hernandez et al., 2012, 2013, 2014). However, in the smart grid system there is also Very Short-Term Load Forecasting (VSTLF) (Hernandez et al., 2012). In any cases, load forecasting is very crucial to all power system operators. In fact, a motivating influence is the market deregulation in most countries. Also, the unbundling of the power sector is a factor. For instance in Nigeria, the formal Power Holding Company of Nigeria (PCHN) is unbundled into three interrelated companies namely Power Generation Company (GenCo), Power Transmission Company (TransCo) and Power Distribution Company (DisCo). Electricity demand forecasting will, therefore, help the power players across classes to manage the power system’s load effectively and efficiently. With load forecasting, the Utilities will especially make essential decisions critical to its operation and planning. The decision can be purchasing and power generational. Others are load switching, infrastructure development, capacity planning, maintenance schedules, energy demand, production adjustment, and contract evaluation (Gamboa, 2017; Kuo & Huang, 2018; Sarabjit & Rupinderjit, 2013).

Certain factors directly influence change in electricity demand of a region. The factors are classified accordingly as economic, demographic and technological. Others are those influenced by policy change and environment (Momani, 2013). For STLF which is the focus of this study, some factors atimes considered are time factors, weather influence and class of the users (Wan, 2014). Accordingly, the time influence can be time of the year, day of the week, and hour of the day. Similarly, there has been observed difference in the demand between weekdays and weekends (Feinberg & Genethliou, 2005; Wan, 2014). More so, there is noticeable change in the demand on holidays, this is lower load than the non-holidays (Wan, 2014). Periodicity of demand occurrence is another factor. STLF usually show noteworthy periodicity. As a matter fact, consumption is a function of daily work and rest period. Thus, it is relevant to be well noted when studying STLF problems. Hourly demands in related days also demonstrate similar patterns (Wan, 2014). Weather factor also greatly influence demand of electricity on the grid. In the tropics where weather situation is average throughout the year, the consumption pattern is obviously different from that of the temperate zones. During the harmattan season and rain season demand patterns exhibit different shapes for obvious reasons. Various weather variables are considered for STLF. Temperature and humidity are the most used for load prediction. Other parameters such as rainfall, wind speed, wind direction, solar irradiations etc. are also considered (Feinberg & Genethliou, 2005; Wan, 2014).

A number of methods have been developed for the prediction of electricity demand from Utilities. A good number of approaches namely, the similar day method, regression analysis, time series, neural networks, expert systems, fuzzy logic, and statistical learning algorithms, are used for STLF (Feinberg & Genethliou, 2005). However, advances in research have led to realisation of more reliable and precise forecasting methods. Some of these methods are classified as classical approach for electricity demand prediction. Some of these methods suffer from imprecise estimation of loads on the power grid. This would, no doubts, cause imbalance in the demand and supply of generated electricity which could lead to supply shortage and wastage (Kuo & Huang, 2018; Wan, 2014). As a result, an accurate method of demand forecasting in the short-term is germane. Machine learning techniques which are capable of learning from data are a more advanced approach to STLF. But, these techniques also have their own problems especially when applied data are a lot plenty while fitting a model for the realisation of its inherent skills on a task. However, the deep learning methods or techniques come with needed rescue. The deep learning techniques are some machine learning algorithms and models that have capabilities to learn tasks and features directly from applied data. Deep learning techniques precisely extract most hidden features underlying and undermining the precision of a model performance on a demand forecasting problem such as the electricity instances. Precise or accurate demand forecasting has advantage of reducing generated costs as well as assists in the reliability of the power sector's responsibilities (Wan, 2014). These, among others, are the promise of deep learning methods. Deep learning is a concept found in the machine learning computing knowledge domain which presents with a lot of research achievements in many areas of computing such as computer vision, signal processing and natural language processing etc. But the techniques have recently been exploited for its applicability to research problems in the power system, especially the electricity demand forecasting.

We therefore investigated some of the deep learning techniques as relevant to the problem area (Adewuyi, Aina, Uzunuigbe, Lawal & Oluwaranti, 2019); and consequently, we exploited only a few models that recent studies have demonstrated to embody positive influence on the STLF research. However, we observed that most of these studies in the problem area were mostly done by applying the temperate climate datasets (Bouktif, Ali, Ali & Mohamed, 2018; Ghullam & Angelos, 2017; Hussein, 2018; Kuo & Huang, 2018; Stuart & Norvig, 2013; Yi, Jie, Yanhua & Caihong, 2013). Most of the approaches implemented were not applicable to nor capable of exploiting feature abstractions characterising a typical tropics dataset needed for attaining a precise STLF model.

In this paper, we developed a STLF model that can reliably and accurately predict day-ahead electricity consumptions. We investigated the process underlying the problems and formulated a precise deep learning model for the process. We, therefore, limited the study to three deep learning techniques well established in the literature (Bengio, 2009; Bouktif et al., 2018; Brownlee, 2018; Chengdong, Zixiang, Dongbin, Jianqiang & Guiqing, 2017; Deng, 2013; Deng & Yu, 2013; Ghullam & Angelos, 2017; Hamedmoghadam, Joorabloo & Jalili, 2018; Hosein & Hosein, 2017; Kuo & Huang, 2018; Schmidhuber & Sepp, 1997; Stuart & Norvig, 2013; Wan, 2014) for modelling electricity demand forecasting. The techniques employed are the Long Short Term Memory (LSTM) network, Convolutional Neural Network (CNN) and Multilayer perceptron (MLP). The data is based on three-year historic electricity data collected from the Transmission Company of Nigeria (TCN) and one year weather data collected from the Nigerian Meteorological Agency (NiMet) for an institutional customers. The datasets are preprocessed for various anomalies, such as inputting the missing values and the models are applied to the datasets with the results being analysed for their training, validation and prediction scores. The results show that upon comparing the three techniques, the LSTM model has an average performance across the training, validation and testing metrics.

2. RELATED WORK

Research studies that are adopting deep learning method have increased in the last decade since Hinton, Osindero & Teh (2006) presented a novel work on deep belief network. A vast majority of works in this domain focused computing areas of acoustic, image, natural language, and signal processing. But, the last few years also witness application of deep neural network approach on power system's datasets especially for load prediction purposes. In the study by (Hossein & Hossein, 2017), a STLF system was developed using Deep Neural Network (DNN) techniques. The studies compared DNN with some traditional methods including moving averages, regression trees and support vector regression.

The DNN methods used are DNN without pretraining (DNN-W), DNN with pre-training using Stacked Autoencoders (DNN-SA), standard RNN and RRN-LSTM. The authors conclude that, in all the DNNs and the baseline techniques, the DNN pretrained with SAE had most stable characteristic when run through both 200 and 400 epochs as it outperformed other methods. The research also used the eventual DNN results to demonstrate dynamic pricing possibility especially on peak load reduction application. Ghullam & Angelos (2017) developed a Feed-Forward DNN and Recurrent DNN models to predict short term electricity load and exploit their applicability. The study introduced utilisation of time/frequency feature extraction procedure initiated by the two models which reveals hidden dominant factors responsible for electricity consumption, as it considers prediction accuracy. Temperate climate datasets were applied. These datasets were analysed on time/frequency domains self-reliantly and the frequency domain components are subsequently transformed back to the time domain which results in capturing of latent features useful for accurate day-ahead electricity demand load measurement. The result of combined features of time/frequency analysis with DNNs enables attainment of higher accuracy. In the work done in (Seunghyoung, Hongseok & Jaekoo, 2017), the researchers identified the need to investigate important aspect of Demand-Side Management (DMS), that is, individual customer loads as a forecasting scheme. As a result, the study adopts DNN as forecasting model for various DSM's loads of a region. The work proposed a STLF framework which is DNN-based. The DNNs are trained in two ways namely: DNN with pretraining scheme using RBM and DNN without pretraining scheme, using ReLu as an activating function, in order to forecast daily loads a day-ahead. The DNNs were compared with two shallow network techniques in its evaluation. The problem of training DNN with customer load was investigated and this revealed that overfitting may occur if the training set is small. The results obtained shown that DNNs performances were better than that of the shallow neural networks. This study also utilises temperate climate datasets. Wan (2014) developed a DNN based load forecaster. He also analyses the critical features underlying load forecasting problems as well as discusses demand forecasting factors to include periodicity, time dependency, holiday effects and weather influences. The researcher presented RBM pretraining and discriminative as pretraining technologies for the problem. He thus utilised 3 years' temperate region datasets, following the algorithms, and compared different neural network models. The result showed that DNNs with pretraining compared favourably and are superior to both DNN without pretraining and single layer ANN. This, thus, supports consideration of a deep learning method to forecasting electricity demand. On the application of a 'community' temperate datasets, the study in (Kuo & Huang, 2018) introduced a precise DNN model for energy load forecasting on short-term horizon using deep learning, a Convolutional Neural Network (CNN) approach. The study compared forecasting results of some AI algorithms which are usually used in load forecasting problem with the CNNs-based DNN, exhibiting great accuracy. Furthermore, the study in (Bouktif et al., 2018) developed an LSTM model for a European electricity consumption data (a typical temperate climate study) using various configurations of the LSTM network for forecast of short to medium term aggregate load forecasting. The study also trained some machine learning algorithms and adopts the best as baseline for comparison with the LSTM. The result shows that with LSTM model, accuracy performance is a lot enhanced unlike an optimized machine learning baseline model. Similarly, on long-term forecasting of electricity demand based on LSTM deep architecture, the study in (Agrawal, Muchahary & Tripathi, 2018) notes that the standard methodology for the LSTM is mainly restricted to electricity demand data characterised by the granularity of a month or a year, leading to very low accuracy load prediction. It therefore developed a method for LTLF having hourly resolution. The model is centred on Recurrent Neural Network (RNN) consisting of the LSTM cells. It also took into consideration the long-term relations in sequence electricity demand data.

In summary, the related work on STLF problem using deep learning approaches have one way or the other demonstrated and shown success on the application of any of the feed-forward or the recurrent deep learning techniques to electricity demand forecasting. However, the data utilised for these studies differs in terms of the climate of the study area. Most of the reviewed works were done for the temperate region. In any situations, there is need to estimate electricity load forecasting problem with interest on the climatic condition of the study area. Weather, being an essential aspect of the climate, appears a big influence on electricity demand, considering how change in weather condition of a place could shape its load profile (Feinberg & Genethliou, 2005; Hernandez et al., 2012; Momani, 2013). Therefore, this study is a typical tropical climate study for an institutional customer type. It is also a typical study for modelling load forecasting problems in the tropics using deep architectures.

3. DEEP LEARNING BASED LOAD FORECAST

3.1. Convolutional neural network

Convolutional Neural Network (CNN) is one of the most popular techniques for deep learning with images and videos. However, its applicability to quantitative data such as the electricity demands is sketchy. The internal representation and architecture of CNN is as in Adewuyi et al., (2019). Furthermore, in order to obtain a richer representation of the applied data, the hidden layer was stacked, so as to obtain multiple feature maps (Hosein & Hosein, 2017). Therefore, in this case, 3 layers of CNN were stacked, with each comprising 128 neurons and 3 kernel size.

3.2. Long Short-Term Memory (LSTM)

LSTM is another deep learning architecture utilised in this study. The introduction to its concepts is documented in our previous work (Adewuyi et al., 2019). However, in this paper, the LSTM deep method was defined as a sequential model characterised by 128 neurons of 3 stacked architecture.

Furthermore, in order to have effective model training, we utilised Truncated backpropagation Through Time (TBPTT) algorithm defined in (Sutskever, 2013) and adjudged as the most practical method for training RNNs models (Brownlee, 2018; Ronald & Jing, 1990)

3.3. Multilayer Perceptron

The Multilayer Perceptron (MLP) is characterised by a system of input, hidden and output layers of neurons. However, unlike an MLP, a perceptron will only have all its input directly connected to its output (Stuart & Norvig, 2013). We, therefore, discuss a single hidden layer MLP for easy digest of its concept.

Assume a situation of n samples of data inputs x_1, x_2,x_3,…, x_{n-2}, x_{n-1}, x_n and corresponding outputs y_1,y_2,y_3,…, y_{n-2}, y_{n-1}, y_n. Therefore, we evaluate the hidden layer input $\hat{y}_h$, expressed as:

$$\hat{y}_h \; = \rho(\textstyle\sum_{k=1}^{n} x_k \times \; w_{kl} - \omega_l) \tag{1}$$

where: w_{kl} – is the weight on data input k and hidden neuron l,
ω_l – is the lth neuron bias, while
$\rho(x)$ – is the sigmoid activation function (Hosein & Hosein, 2017; Stuart & Norvig, 2013) defines as:

$$\rho(x) = \frac{1}{1+e^{-x}} \tag{2}$$

Following this is the output estimation of the MLP. We defined this as:

$$\hat{y}_j \; = \rho\left(\textstyle\sum_{l=1}^{m} x_{lj} \times \; w_{lj} - \omega_j\right) \tag{3}$$

where: m – is the total inputs neurons, j, at the output layer.

Consequently, an update of the weight follows. This is necessary in order to reduce error rate (Brownlee, 2018; Hosein & Hosein, 2017; Stuart & Norvig, 2013) estimated as:

$$e = \; \hat{y}_j - y_j \tag{4}$$

So, this study defined MLP also as sequential model also with 128 neurons and 3 stacked layers.

4. ANALYSIS

4.1. Data Description

The applied datasets primarily consists of 25,751 samples of 21 features representing three years institutional electricity consumptions from the national grid and some weather data. The power data were recorded at hourly intervals throughout the years and it represents a typical tropics electricity demand. The power data was collected at the Transmission Company of Nigeria (TCN) 132/33kV, Ile-Ife while the weather parameters were taken at the Nigerian Meteorological Agency (NiMet) situates at Ido-Osun Aerodrome, Osogbo, Nigeria. These datasets were initially split into two parts, 80% training set and 20% test set. The test set was consequently broken into 60% validation set and 40% test set. As can be inferred, the datasets are characterised by both electrical and non-electrical features. The electrical features include the load taken at each hour of the day and other lagged loads such as the previous one hour, previous two hours, previous day same hour, previous day previous hour, previous day previous two hours, previous two days same hour, previous two days previous hour, previous two days previous two hours, previous week same hour, average of past twenty-four hours, average of past seven days, day of the week, weekend-day and holiday. Understanding what type of data our problem represents, suggests that we carefully culled out ‘holiday’ consumptions from our dataset as stated herein. The non-electrical data features include weather parameters such as relative humidity, dry bulb and wet bulb. We also included calendar features such as the actual date load was recorded, time load was recorded and date/time load was taken in a particular day.

4.2. Modelling Method

As a data preprocessing strategy, our data is cleansed by removing noisy data such as an instance of incorrect date value. Missing data was handled by *imputing* values using the mean strategy (Swalin, 2018). Furthermore, we prepare the data by scaling the numeric data and transforming the categories. This is with a view to improving the stability of the network and modelling performance. We, therefore, standardised our data to have zero mean and unit variance. This will enable the model to be more robust to new data. For categorical data, we adopt a one-hot encoding approach as in (Swalin, 2018). This prevents ordering, leading to model poor performances.

As a compilation strategy, we carefully specified some parameters for the network training purposes. We simply set Adam optimiser to the following configuration: exponential decay rate = $1e-3$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ which is in line with the algorithm in (Swalin, 2018) and specified below:

```
Require: $\alpha$: Stepwise
Require: $\beta_1\beta_2 \in (0,1)$: Exponential decay rate for the moment estimates
Require: $f(\theta)$: Stochastic objective function with parameter
Require: $\theta_0$: Initial parameter vector $m_0 \leftarrow 0$: Initialises first moment vector
    $v_1 \leftarrow 0$: Initialises second moment vector
    $t \leftarrow 0$: Initialises timestep
while $\theta_0$ unconverges repeat
    $t \leftarrow t + 1$
    $g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$: Gets gradient w.r.t stochastic objective at time-step $t$.
    $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot g_t$: Updates biased first moment estimate.
    $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1-\beta_2) \cdot g_t^2$: Update biased second raw moment estimate.
    $\widehat{m_t} \leftarrow m_t/(1-\beta_1^t)$: Computes the bias-corrected first raw moment estimate.
    $\widehat{v_t} \leftarrow v_t/(1-\beta_2^t)$: Computes the bias-corrected second raw moment estimate.
    $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \widehat{m_t}/(\sqrt{\widehat{v_t}}+\epsilon$: Updates parameters.
end while
    return $\theta_t$: Resulting parameters,
where $\leftarrow$ means equal to.
```

Furthermore, in order to configure the models loss function, we considered the Mean Squared Error (MSE), which is defined as below:

$$MSE = \epsilon[(y_m - \bar{y}_m)^2] \tag{5}$$

As a model fitting approach, we specified a matrix of input features and corresponding output patterns. Our model is, therefore, fit by means of the Truncated Backpropagation Through Time (TBPTT) algorithm as in (Sutskever, 2013).

4.3. Evaluation Method

As an evaluation method, the developed model was based on the three performance evaluation metrics in (Adewuyi et al., 2019).

5. RESULTS

The three models were compared against one another based on the set performance evaluation criteria. To begin with, the learning curves of the developed LSTM model was visualised to see how close or far it is from the regression line. This was done for an experimental cycle of the training set spanning 100 epochs. The same was done for the candidate models CNN and MLP, also with a view to observing how close they are to an ideal solution. This is shown in Fig. 1, Fig. 2 and Fig. 3.

Also, the validation and prediction tests of the developed LSTM model as well as its candidates were carried out. This was with a view to determining how capable they are at forecasting electricity demand from the grid when fed with unseen load data. This validation and test methods follow that a performance analysis of the three techniques were done by comparing the candidate models with the LSTM model using their MSE, RMSE and MAE scores when experimented at different epochs as shown in Tab. 1. Prediction scores of the three applied techniques show that our LSTM model out-performed the CNN and MLP, its candidates, across metrics and epochs. The models were shown our datasets during the training, evaluation and testing phases, each at 40, 60, 80 and 100 epochs to observe their prediction accuracy. Although, we noticed an instance of MLP superiority over LSTM and CNN, but this occurred at an early stage of the experiment, 40 'epoch', precisely. This would amount to nothing significant because a further exposure of our datasets shown supremacy of the LSTM model over others for our problem.

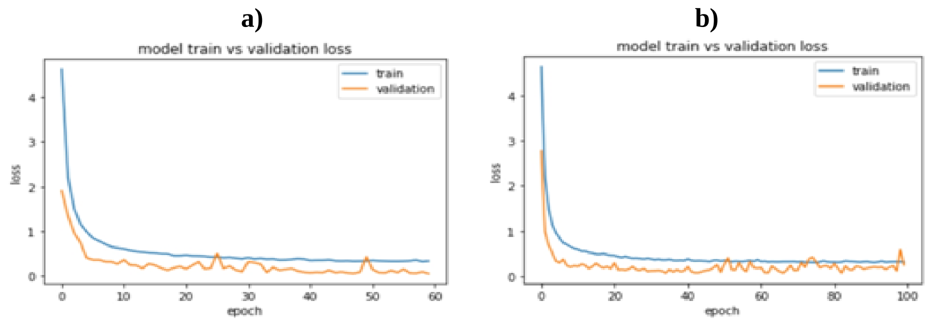


Fig. 1. LSTM model learning losses at 60 and 100 epochs

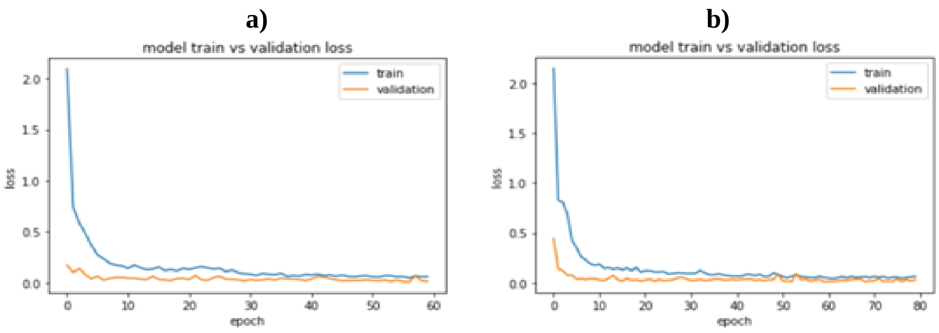


Fig. 2. CNN model learning losses at 60 and 80 epochs

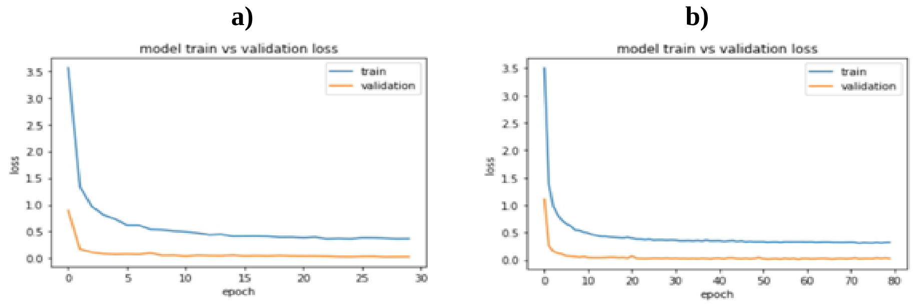
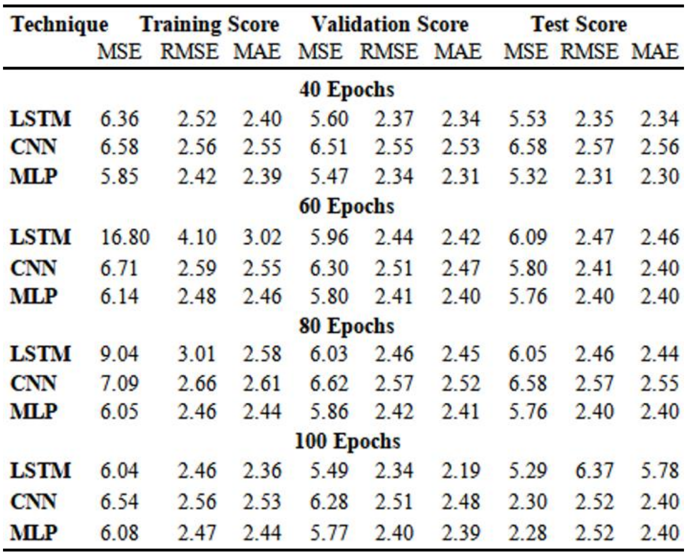


Fig. 3. MLP model learning losses at 30 and 80 epochs

Tab. 1. Comparison analysis of the deep learning techniques performances



Technique	Training Score			Validation Score			Test Score		
	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE
				40 Epochs					
LSTM	6.36	2.52	2.40	5.60	2.37	2.34	5.53	2.35	2.34
CNN	6.58	2.56	2.55	6.51	2.55	2.53	6.58	2.57	2.56
MLP	5.85	2.42	2.39	5.47	2.34	2.31	5.32	2.31	2.30
				60 Epochs					
LSTM	16.80	4.10	3.02	5.96	2.44	2.42	6.09	2.47	2.46
CNN	6.71	2.59	2.55	6.30	2.51	2.47	5.80	2.41	2.40
MLP	6.14	2.48	2.46	5.80	2.41	2.40	5.76	2.40	2.40
				80 Epochs					
LSTM	9.04	3.01	2.58	6.03	2.46	2.45	6.05	2.46	2.44
CNN	7.09	2.66	2.61	6.62	2.57	2.52	6.58	2.57	2.55
MLP	6.05	2.46	2.44	5.86	2.42	2.41	5.76	2.40	2.40
				100 Epochs					
LSTM	6.04	2.46	2.36	5.49	2.34	2.19	5.29	6.37	5.78
CNN	6.54	2.56	2.53	6.28	2.51	2.48	2.30	2.52	2.40
MLP	6.08	2.47	2.44	5.77	2.40	2.39	2.28	2.52	2.40

6. CONCLUSIONS

In this paper, we demonstrate an application of deep learning techniques for electricity demand forecasting. We considered three deep architectures (LSTM, CNN and MLP) for our study. The applied datasets, which is a typical University electricity consumption, underpins the concepts of weather influence on loads especially in the tropics. These three techniques were compared to see which deep method would perform best. Thus, we found out that the LSTM model, of the three techniques out-performed others. We, therefore, establish that deep learning approach is a suitable approach to solving the problem of electricity demand forecasting.

REFERENCES

Adewuyi, S., Aina, S., Uzunuigbe, M., Lawal, A., & Oluwaranti, A. (2019). An Overview of Deep Learning Techniques for Short-Term Electricity Load Forecasting. *Applied Computer Science*, *15*(4), 75–92. doi: 10.23743/acs-2019-31

Agrawal, R. K., Muchahary, F., & Tripathi, M. M. (2018). Long term load forecasting with hourly predictions based on long-short-term-memory networks. In *2018 IEEE Texas Power and Energy Conference (TPEC)* (pp. 1–6). College Station, TX.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundation and Trends in Machine Learning, 2*(1), 1–127.

Bouktif, S., Ali, F., Ali, O., & Mohamed, A. S. (2018). Optimal Deep Learning LSTM Model for Electric Load Forecasting using Feature Selection and Genetic Algorithm: Comparison with Machine Learning Approaches. *Energies*, *11*, 1636–1656.

Brownlee, J. (2018). *Deep learning for time series forecasting: Predicting the future with MLPs, CNNs and LSTMs in Python*. V1.2 ed. M. L. Mastery.

Chengdong, L., Zixiang, D., Dongbin, Z., Jianqiang, Y., & Guiqing, Z. (2017). Building energy Consumption prediction: An extreme deep learning approach. *Energies, 10*(10), 1525–1545.

Deng, L. (2013). A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Transactions on Signal and Information Processing*, *3*(2). doi:10.1017/ATSIP.

Deng, L., & Yu, D. (2013). Deep learning: Methods and Applications. *Foundations and Trends in Signal Processing, 7*(3-4), 197–387.

Feinberg, E. A., & Genethliou, D. (2005). Load forecasting. In J. H. Chow, F. F. Wu, J. Momoh (Eds.), *Applied Mathematics for Restructured Electric Power Systems*. *Power Electronics and Power Systems, Springer* (pp. 269–285). Boston, MA.

Gamboa, J. (2017). Deep learning for time-series Analysis. *arXiv: 1701.01887[cs. LG]*.

Ghullam, M. U., & Angelos, K. M. (2017). Short-term power load forecasting using deep neural networks. *ICNC*, *10*(1109), 594–598.

Hamedmoghadam, H., Joorabloo, N., & Jalili, M. (2018). Australia's long-term electricity demand forecasting using deep neural networks, *arXiv:1801.02148 [cs.NE]*.

Hernandez, L., Baladron, C., Aquiar, J. M., Calavia, L., Carro, B., Sánchez-Esguevillas, A., Cook, D. J., Chinarro, D., & Gomez, J. (2012). A study of relationship between weather variables and electric power demand inside a smart grid/ smart world. *MDPI Sensors*, *22*(9), 11571–11591. doi:10.3390/s120911571

Hernandez, L., Baladron, C., Aquiar, J. M., Calavia, L., Carro, B., Sánchez-Esguevillas, A., Cook, D. J., Chinarro, D., & Gomez, J. (2013). Short-term load forecasting for micro-grids based on artificial neural networks. *MDPI Sensors*, *6*(3), 1385–1408.

Hernandez, L., Baladron, C., Aquiar, J. M., Calavia, L., Carro, B., Sánchez-Esguevillas, A., Perez, F., Fernández, A., & Lloret, J. (2014). Artificial neural network for short-term load forecasting in distribution systems. *MDPI Energies, 7*(3), 1576–1598.

Hosein, S., & Hosein, P. (2017). Load forecasting using deep neural networks. In *Proceedings of the Power and Energy Society Conference on Innovative Smart Grid Technologies* (pp. 1–5). IEEE.

Hussein, A. (2018). *Deep Learning Based Approaches for Imitation Learning* (Doctoral dissertation). Robert Gordon University, Aberden, Scotland.

International Energy Agency (IEA) Publications and data (n.n.). Retrieved August 12, 2018 from https://www.iea.org

Kuo, P., & Huang, C. (2018). A high-precision artificial neural networks model for short-term energy load management. *Energy, 11*(1), 213– 226.

Momani, M. A. (2013). Factors Affecting Electricity Demand in Jordan. *Energy and Power Engineering*, *5*, 50–58.

Ronald, J. W., & Jing, P. (1990). An Efficient Gradient-Based Algorithm for On-Line Training of Recurrent Network Trajectories. *Neural Computation, 2,* 490–501.

Sarabjit, S., & Rupinderjit, S. (2013). ARIMA Based Short Term Load Forecasting for Punjab Region. *IJSR, 4*(6), 1919–1822.

Schmidhuber, J., & Sepp, H. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Seunghyoung, R., Hongseok, K., & Jaekoo, N. (2017). Deep neural network based demand side short term load forecasting. *Energies MDPI, 10*(1), 3–23.

Stuart, R., & Norvig, P. (2013). *Artificial Intelligence A modern Approach*. Second ed. Prentice Hall.

Sutskever, I. (2013). *Training Rucurrent Neural Net-works* (Doctoral dissertation). Computer Science, University of Toronto, Toronto.

Swalin, A. (2018). *How to Handle Missing Data.* Towards Data Science. https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4 on 18/01/19.

Wan, H. (2014). Deep Neural Network Based Load Forecast. *Computer Modelling and New Technologies*, *18*(3), 258–262.

Yi, Y., Jie, W., Yanhua, C., & Caihong, L. (2013). A New Strategy for Short-Term Load Forecasting. *Hindawi*, *2013,* 208964. doi:10.1155/2013/208964

Applied Computer Science, vol. 16, no. 1, pp. 18–32
doi:10.23743/acs-2020-02

Submitted: 2020-03-13
Revised: 2020-03-20
Accepted: 2020-03-23

similarity metrics, recruitment, workers, college graduates

Mario BELLO[*], *Alejandra LUNA*[*], *Edmundo BONILLA*[*], *Crispin HERNANDEZ*[*], *Blanca PEDROZA*[*], *Alberto PORTILLA*[**]

A NOVEL PROFILE'S SELECTION ALGORITHM USING AI

Abstract

In order to better understand the job requirements, recruitment processes, and hiring processes it is needed to know the people skills. For a recruiter this entails analyzing and comparing the curricula of each available candidate and determining the most appropriate candidate that the activities that are required by the position. This process must be carried in the shortest length of time possible. In this paper, an algorithm is proposed to identify those candidates, either workers or college graduates.

1. INTRODUCTION

The recruitment and selection is one of the most ancient areas of applied psychology, besides, it is one of the most important domains in talent management and human resources (Derous & Fruyt, 2016). The evolution of the labor market has caused the traditional recruitment process to not be enough. Nowadays, the internet has introduced new methods to carry the recruitment process, starting with a new form to generate or creating a resume and the way in which it is distributed to the companies (Kesler, Béchet, Roche, Torres-Moreno & El-Bèze, 2012). For this reason, organizations have started to use different technological platforms to lure personnel as part of the electronic recruitment (Esch & Mente, 2018). Some organizations have even started to adopt artificial intelligence methodologies in their recruitment processes (Esch, Black & Ferolie, 2019).

[*] Tecnológico Nacional de México/Instituto Tecnológico de Apizaco, 90300, Carretera Apizaco Tzompantepec, Esquina Av., Instituto Tecnológico S/N, Apizaco, Tlaxcala, México, edbonn@walla.co.il

[**] Smartsoft America Business Applications S.A. de C.V., 90806, Adolfo López Mateos S/N, Texcacoac, Chiautempan, Tlaxcala, México, aportilla@smartsoftamerica.com.mx

In order to help the recruiter in the search of the most suited profile, a profile selection algorithm is proposed in this paper. We use a methodology that employs two search criteria of profiles, guaranteeing that only the profiles that meet one of the criteria are analyzed. The profiles belong to real people. These were taken from web sites such as Indeed.com and Mexico's employment web page (www.empleo.gob.mx). At the same time, the methodology includes a pre-processing stage to standardize the profiles as well as eight similarity metrics (Cosine, Euclidean, Levenshtein, Dice, N-grams, Jaccard, Fuzzy distance and Q-grams), in charge of finding the similarity degree between the profile and the employment vacancy. This algorithm was developed for the platform I'm talenty, owned by Smartsoft America Business Applications. The purpose of the platform is early linking in which students, companies and educational institutions interact (I'm Talenty, 2019). Smartsoft is a TI company that develops innovative solutions for the Mexican market.

2. RELATED PAPERS

In the metric similarity field, an adjustable approach of object parameters to predict unknown data in soft incomplete fuzzy sets was proposed, this is based on the similarity metrics of fuzzy sets (Liu, Qin, Rao & Mahamadu, 2017). Kerzendorf (2019) presented the application of computational linguistic techniques into the literature within the field of astronomy, which is a result of the recommendation of scientific articles or reference texts. Another advance in the document grouping was originated with the implementation of the N-grams technique and the enhance squared cosine similarity (Bisandu, Prasad & Liman, 2018), the methodology consisted in preprocessing new scientifically articles. (Cheatham & Hitzler, 2013) It was focused on the ontological alignment systems that implement chain similarity metrics. A basic system was also developed to automatically select a similarity metric of each chain set for a pair of ontologies in execution time, based on the characteristics of the ontologies to be aligned.

In the work matching field, a deep neural network with imbedded layer model was implemented to predict the future professional details of an employee (Deng, Lei, Li & Lin, 2018), based on the data at an online resume. It was proved that this model was more effective and appropriated than methodologist such as Random Forest, XGB or a deep neural network. Another approach for the matching between the offer and the demand was proposed on ESCO ontology (Shakya & Paudel, 2019), which is a multilingual classification of abilities, competences, qualification, and European occupations. It uses the similarity score, which is a measure to show how alike two sets of data are. Another advance was done when a presentation of the approach for the alignment of online profiles and job announcements mixing themes of a thesauri with Levenshtein distance, the Dice's coefficient and the Okapi BM25 measure (González-Eras & Aguilar, 2019).

3. METHODOLOGY

In this work, we propose to use similarity metrics to analyze two elements and to find a similarity degree. The elements to be analyzed are:

- College graduate's profiles,
- Worker profiles,
- Job offers.

The proposed algorithm is illustrated in Figure 1, all the elements that comprise it are observed.

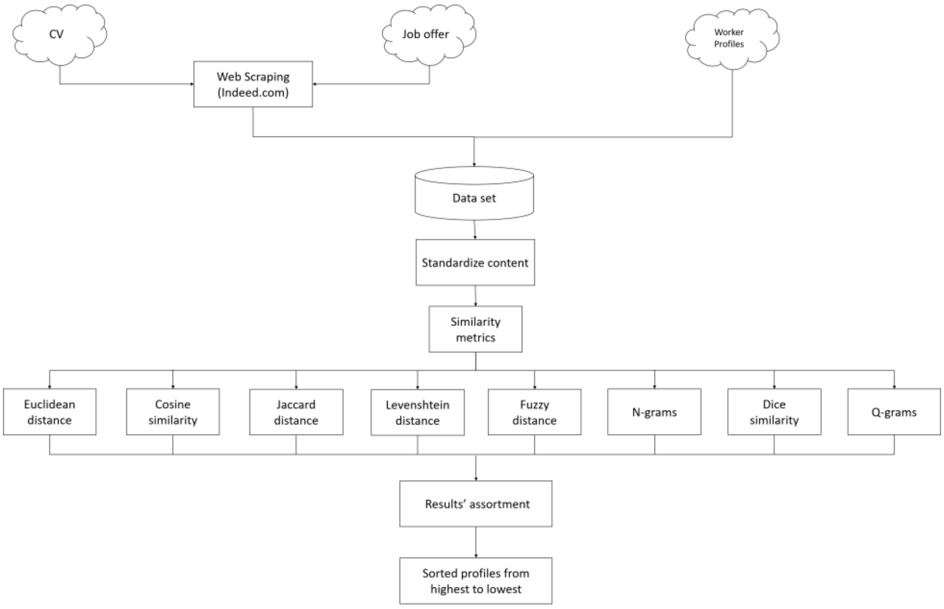


Fig. 1. Profiles selector algorithm

1. Data set – Is the type SQL database, where the profiles and the job offers are kept.
2. Standardize content – Before searching the similarity degree between the profiles and/or job offers the content of this must be standardized.
3. Similarity metrics – Once the elements to be analyzed have been standardized, the search for a similarity degree follows.
4. Results' assortment – Once all the elements had been analyzed either finding a college graduated for a job offer or searching for a worker profile, the result are ordered.
5. Ordered profiles – At last, it gives back the user the list of sorted profiles.

3.1. Data set

The profiles of college graduates and the job offers were obtained by using the Web Scraping technique on Indeed.com. The workers' profiles were obtained manually from Mexico's employment portal (www.empleo.gob.mx), due to the lower amount in this kind of profiles. We should keep in mind that the information gathered is in Spanish and matches real profiles and job offers.

Because of this, the extracted profiles from Indeed.com and Mexico's employment portal are not structured, for every user is free to describe his/her profile as he/she sees fit. Before saving the profiles in the database, it was necessary to sort them by type. The content of workers' profiles and college graduates was structured in fields such as:

- Id – It corresponds to an identification number and it was assigned in increasing order and without repetition.
- Career – It corresponds to the college graduate's career or the employment of the worker's profile. This field uses as a criteria of profile search.
- Specialty – It corresponds to the specialty that the candidate or worker have (in case it has it). This field is used as a profile search criterion.
- State – It indicates the state or city in which the workers or the college graduate lives.
- Description – It corresponds to the previous abilities knowledge and/or experience that are contained by the profile. This last field is taken into account to search for the similarity degree.

The content of job offers and workers profiles was structured in a similar way as the college profiles, this was because only one field was added:

- University – It corresponds to the university of the college graduate. For the worker's profile, it corresponds to their education level.

The database contains 154 records, divided on the college graduated profiles, the workers and the job offers.

3.2. Content standardization

Before looking for a candidate meeting a specific job offer, it is necessary to standardize the content of the elements under consideration.

This process implies the elimination of grammatical elements in the Spanish language. Some elements are:

- Specific or not specific articles – a/an/the,
- Possessives – Mine/your/his/ours,
- Demonstratives – This/that/these/those.

Other elements to eliminate are punctuation marks (dot, colon, semi-colon, quotation marks, etc.). The same rule applies for special characters (@, $, *, <, etc.).

At last, the rest of the content of the elements are switched to lower case. This is due to the possibility of a word being written in a different way. For instance, the programming language Java can be written as java, JAVA, etc. For a system engineer, this has the same meaning, however, for similarity metrics it implies a minimum difference. All the elements described previously are found in a dictionary, allowing you to continue adding more grammatical elements that can be discarded from the profiles.

The standardization process is carried in order to minimize orthographic mistakes besides eliminating those words that do not have useful information and interfere with the similarity metrics analysis. At the same time, this process is executed whenever it is necessary to look for a candidate for a work position, and it is applied to the profiles and offer.

3.3. Similarity metrics

To know the similarity among profiles and a job offer, we propose to use similarity metrics. A similarity metric reflects the closeness between two objects, it must correspond to the characteristics that are thought to be integrated in the data groups. In this document eight similarity metrics are used.

Euclidean distance. It is a standard metric for geometrical issues. It is the ordinary distance between two points, which can be easily measured with a ruler in a dimensional or tridimensional space. It is widely used in problems clustering, even in text clustering (Huang, 2018). The Euclidean distance of two documents is defined by equation (1).

$$D_E\left(\overrightarrow{t_1}, \overrightarrow{t_2}\right) = \left(\sum_{t=1}^{m}\left|W_{t,1} - W_{t,2}\right|^2\right)^{1/2} \tag{1}$$

Given two documents (t_1, t_2) represented by their vector terms $\overrightarrow{t_1}$ and $\overrightarrow{t_2}$, further it´s term sets $T = t_1, \dots, t_m$.

Cosine similarity. This metric is based in angles and orientation between two vectors discarding their longitude, which means it is the same that the cosine of the angle between vectors (Sandhya, Lalitha, Govardhan & Anuradha, 2008). In equation (2), the cosine similarity is represented.

$$SIM_{c(\overrightarrow{t_1},\overrightarrow{t_2})} = \frac{\overrightarrow{t_1} \times \overrightarrow{t_2}}{\|\overrightarrow{t_1}\| \|\overrightarrow{t_2}\|} \tag{2}$$

In the equation (2), $\overrightarrow{t_1}$ and $\overrightarrow{t_2}$ are considered to be m-dimensional vector on the terms set $T = t_1, \dots, t_m$. Each dimension represents a term in the document, which is not negative. As a result, the value given to that metric is delimited in the interval [0, 1].

Jaccard's distance. It measures the similarity of two elements in a way that the intersection of the elements is divided between the data element union (Guo, Jerbi & O'Mahony, 2014). This metric is represented in equation (3).

$$SIM_J\left(\overrightarrow{t_1}, \overrightarrow{t_2}\right) = \frac{\overrightarrow{t_1} \cdot \overrightarrow{t_2}}{\left|\overrightarrow{t_1}\right|^2 + \left|\overrightarrow{t_2}\right|^2 - \overrightarrow{t_1} \cdot \overrightarrow{t_2}} \qquad (3)$$

For the documents t_1 and t_2, the Jaccard coefficient compares the sum of the terms that appeared in any of the documents but that are not shared. The result of this metric is in the interval [0, 1].

Levenshtein distance. It is a proximity measurement between two strings that applies mainly to the sequence comparison in the linguistic domain, like detecting plagiarism and speech recognition (Behara, Bhaskar & Chung, 2018). Levenshtein distance calculates the less expensive set of intersections, eliminations, or substitutions that are required to transform a chain into another. This metric is represented in equation (4).

$$D_L(t_1, t_2) = min_s\left(\sum_{k=0}^{k=s} \beta_k\right) \qquad (4)$$

It defines $S = S_0, S_1, \dots, S_k, \dots, S_s$ as the sequence of edition operations to transform the string t_1 to t_2, after the associated cost to each edition operation as $\beta_0, \beta_1, \dots, \beta_k, \dots, \beta_s$.

Fuzzy distance. These distances are used to compare different objects. Their definition is based in proximity, fuzzy set operation, etc. That makes different property prepositions in the similarity measures (Baccour, Alimi & John, 2014). The measurement based in the fuzzy union and intersection operations is defined on the equation (5).

$$M_{A,B} = \frac{\sum i(a_i \wedge b_i)}{\sum i(a_i \vee b_i)} \qquad (5)$$

With the use of the equation (5), the similarity degree is obtained $M_{A,B}$ from fuzzy sets A y B (Pappis and Karacapilidis, 1993). Another metric of fuzzy similarity is the metric based on the difference and the sum of membership degrees. The equation (6) shows said metric.

$$S_{A,B} = 1 - \frac{\sum i|a_i - b_i|}{\sum i(a_i + b_i)} \qquad (6)$$

In equation (6), the similarity degree $S_{A,B}$ is obtained from the fuzzy sets A y B.

Q-grams similarity. This metric divides a string into substrings of length *q*. The reason behind q-grams is that the characters sequence is more important than the character by themselves. The q-grams similarity is represented on equation (7).

$$SIM_Q(t_1, t_2) = 1 - \frac{\sum_{i=1}^{n} |match\ (q_i, Q_{t1}) - match(q_i, Q_{t2})|}{|Q_{t1}| + |Q_{t2}|} \tag{7}$$

The q-grams for a string t is obtained as a longitude vector space q over the string. We should also consider the longitude substrings q-1 and recognize the prefixes and suffixes of the string, called filler characters (#, %, $) are added at the beginning and at the end of the string (Gali, Mariescu-Istodor, Hostettler & Fränti, 2019).

Dice´s similarity. This similarity metric is based on the absence and presence of words in two documents t_1 and t_2. Said metric is represented on equation (8).

$$SIM_d(t_1, t_2) = \frac{2|t_1 \cdot t_2|}{|t_1|^2 + |t_2|^2} \tag{8}$$

The main aspect of this metric is that it multiplies by two the total number of terms in two documents (Dice, 1945).

N-grams' similarity. It consists of the generalization of the longest common subsequence concept to include n-grams, by only including uni-grams (Kondrak, 2005). This metric is shown on equation (9).

$$SIM_n(\Gamma_{i,j}^{n}) = \frac{1}{n} \sum_{u=1}^{n} S_1(X_i + u, Y_i + u) \tag{9}$$

This metric formulates the similarity of n-grams as a function sn, where n is a fix parameter, while S_1 is equivalent to the function of the uni-grams similarity.

3.4. Sort the profiles and return to the user

When the similarity metrics finishes the evaluation of all the elements, what follows is to sort the results. The assortment of the results is done by using a Merge Sort algorithm, which carries a stable execution and stands out from other sorting algorithms (Sedgewick & Wayne, 2011). John Von Neumann developed said algorithm in 1945, and it is based on the divide and conquer technique. In broad sense the algorithm works as following:

- If the longitude of the list is 1 or 0, then it is sorted,
- The list is divided on two sorted lists of almost the same size,
- Each sub list is sorted repetitively by using the Merge Sort algorithm,
- The bub list that had been sorted are incorporated in one list.

The algorithm takes into account the mean of all the generated results by the similarity metrics of each analyzed element so that the candidate with the most similarity shows up at the beginning of the list, and the candidate with the less similarity, at the bottom. This list is finally sent back to the user for the decision-making.

4. TEST AND RESULTS

Fort the test of the proposed methodology, a SQL database was used, with a total of 154 registers among worker profiles, college graduates' profiles and job offers, taking into account that all the obtained data is from real individuals.

The search for both profiles can be done by using two criteria, by using the field Career or the field Specialty. For instance, if a worker is required, the search criteria will include what one or another field contains in the job offer, to later be able to search on the available profiles to proceed with the selection of profiles that matches the specific fields. The selected profiles are the analyzed using the metrics, in this way the qualification of all the profiles by the metrics is avoided. For the tests, there are two types of profiles to analyze, the profiles of workers and the profiles of college graduates, which are detailed below.

4.1. Workers' search

The process of searching workers is as follows, a driver vacancy with several abilities is taken as a base: "loading and unloading maneuvers, knowledge of traffic regulations, drive different transport units and 3 years' experience driving 3½ tons units". For this profile, the algorithm identifies eight candidates for the vacancy. In Table 1, only the best five profiles are shown, each profile is evaluated by the eight metrics generating eight results in a range [0,1]. If the value shown by the metrics is 1, it means that both the job offer and worker profile are identical, however if the value is 0, then they are different.

Tab. 1. Assessment of better qualified profiles for the driver vacancy

Profile Id	D_E	SIM_C	SIM_j	D_L	Fuzzy distance	SIM_Q	SIM_D	SIM_N
17	0.80	0.93	0.69	0.30	0.44	0.17	0.19	0.28
19	0.78	0.90	0.71	0.37	0.5	0.24	0.23	0.35
20	0.81	0.93	0.83	0.27	0.48	0.33	0.37	0.26
22	0.76	0.92	0.69	0.36	0.46	0.11	0.12	0.35
23	0.78	0.96	0.67	0.28	0.43	0.13	0.15	0.27

The algorithm determines that the No. 20 profile is the best candidate, this by obtaining the mean of the eight similarity metrics. In Figure 2, shows the similarity of each metric for the best profile found. It is worth mentioning that these results are the product of standardization of the content of elements.

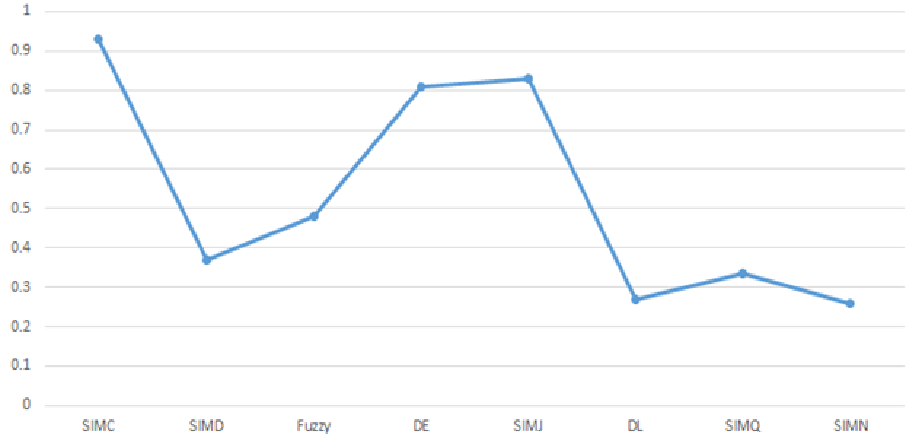


Fig. 2. Evaluation of each similarity metric for profile No. 20

Figure 3 shows how the best five profiles found are qualified by the similarity metrics.

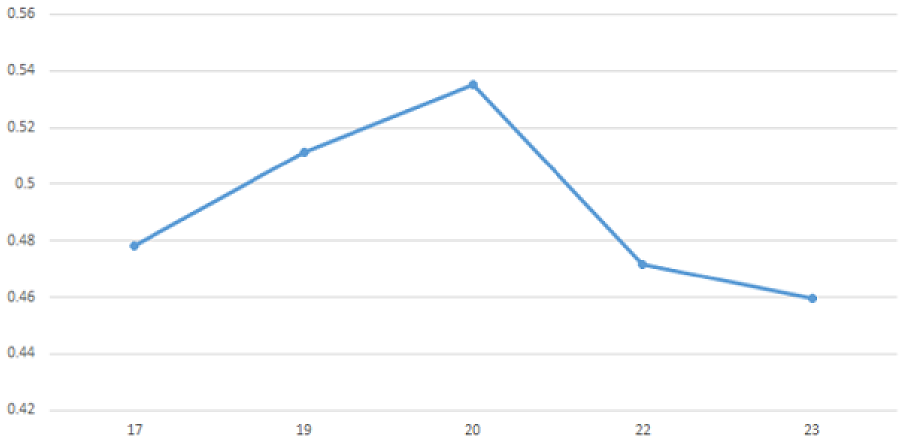


Fig. 3. Better qualified profiles for driver

It is confirmed in Figure 3, that the profile better qualified is profile No. 20, which has abilities such as: “1 year of experience driving 3 ½ ton units, current license, knowledge of traffic regulations, knowledge in the San Luis Potosí area and its surroundings". The worker search process, standardization of the content elements, analysis of the elements by the similarity metrics and presentation of the candidates to the user were done in 3 seconds. In Figure 4, a small comparison of results it is shown when the contents of the elements had been standardized and when they had not been standardized and processed directly.

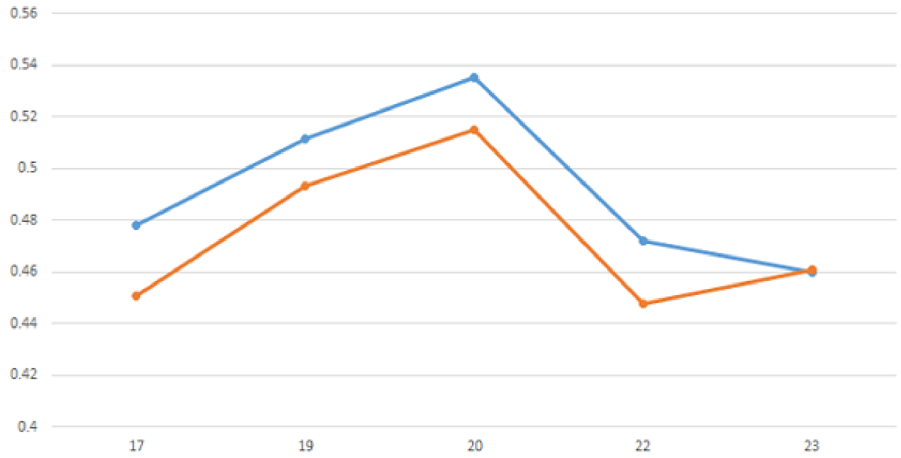


Fig. 4. Better five profiles result comparison with and without standardization

In Figure 4, the blue line belongs to the results after the standardization of the content, and the orange one belongs to the results obtained without standardization. Even though, in both cases, the best profile is profile No. 20. It is confirmed that by standardizing the similarity degree, it is higher. This is because the elements that are omitted on the profiles and job offers are more common, and this interferes with the metric analysis. In a way that these results improve when only the words that provide information or describe the person are taken into account.

4.2. College profiles search

The process of searching college graduated profiles is as follows, based on a job offer that requires a computer systems engineer with skills such as: "experience not required, availability to change residence, knowledge of object oriented programming with java, C#, Python, relational databases with MySQL, SQL server, Oracle or similar, basic knowledge of web design, HTML, CSS, JavaScript, English language skills". The algorithm identifies seven candidates. In Table 2, only shows the best five candidates. In a same way as in the search of a worker, if the value shown by metrics is 1, it means that the job offer and college graduate profile are identical, but if the value is 0, then they are different.

Tab. 2. Better qualified profiles for the computational systems offer

Profile Id	D_E	SIM_C	SIM_j	D_L	Fuzzy distance	SIM_Q	SIM_D	SIM_N
3	0.82	0.95	0.82	0.24	0.35	0.22	0.26	0.21
5	0.87	0.97	0.83	0.23	0.36	0.12	0.17	0.20
6	0.85	0.98	0.77	0.28	0.40	0.07	0.06	0.25
7	0.83	0.98	0.77	0.27	0.41	0.06	0.07	0.26
9	0.84	0.97	0.72	0.26	0.39	0.02	0.02	0.23

The algorithm determines that the best profile is No. 3. This is accomplished by obtaining the mean of all metrics. In Figure 5, the similarity degree for every metric is shown.

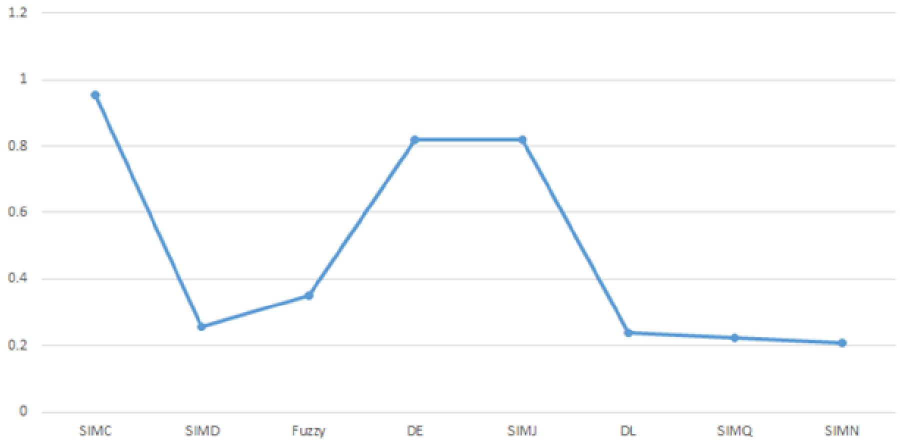


Fig. 5. Similarity metrics evaluation for profile No. 3

The Figure 6 shows the five best profiles evaluated by the similarity metrics.

It is confirmed that the best qualified profile by the metrics is No. 3, which has the following abilities: “knowledge about Microsoft Office, TOEIC certification, knowledge of databases with SQL, SQL server, MySQL, web page design with PHP and Java, maintenance of servers and knowledge in computer networks and 4 years’ experience”. The complete search of candidate process for this scenery was done in 3 seconds.

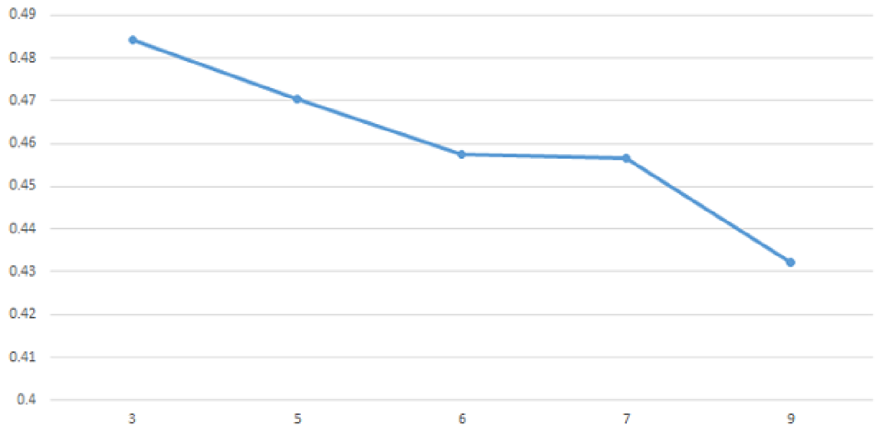


Fig. 6. Better qualified profiles for the computational systems engineer offer

In Figure 7, a comparative of the result when the process has been and has not been standardized is shown.

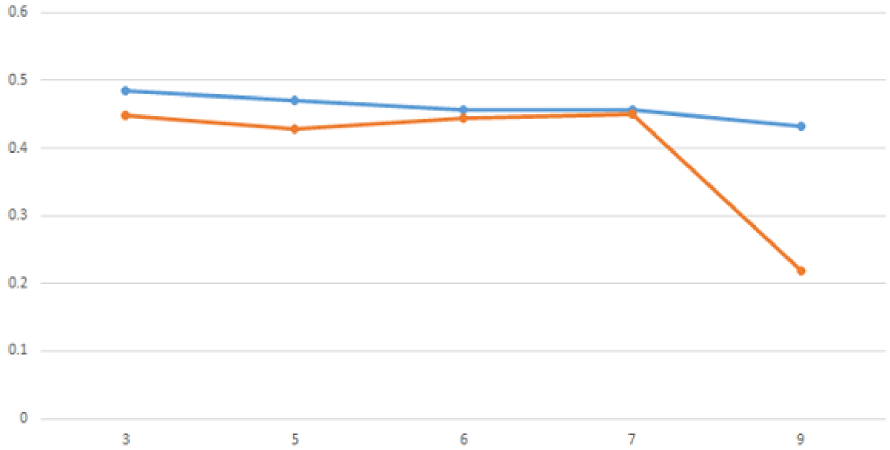


Fig. 7. Better five profiles result comparison with and without standardization

As the previous case, the blue line belongs to the result for the standardized profile, and the orange line belongs to the profile processing without standardizing. The same pattern, shown in the previous case, can be observed, and the similarity degree is higher when the elements have been standardized.

It also worth mentioning that the selection algorithm is not limited to the profile analysis in Spanish. It also works in English, however, a dictionary for that language has to be designed to be able to standardize the profiles. Figure 8 shows the evaluation of the algorithm for a robotics researcher vacancy. The analyzed profiles are in English.

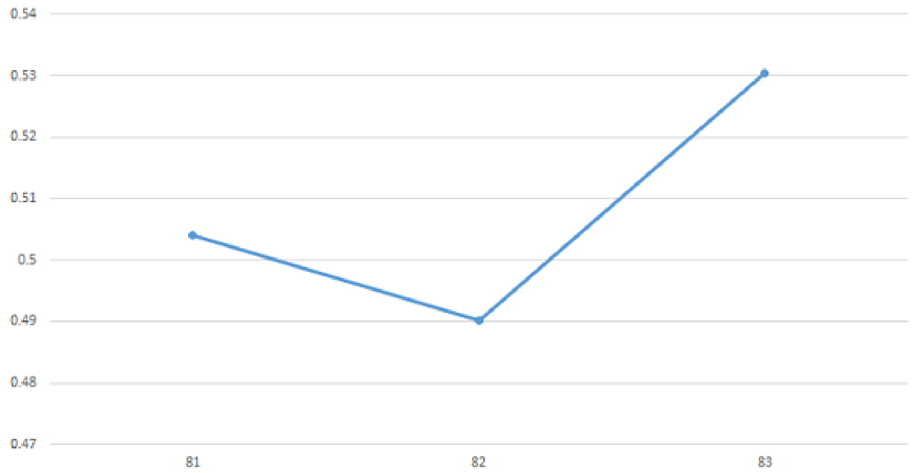


Fig. 8. English language evaluated profiles

The algorithm finds three candidates and the best evaluated for this is profile 83, which corresponds to a researcher.

The profile selection algorithm cannot be compared directly to other methodologies, because it carries different techniques in order to be able to align the profiles with job offers. Table 3 shows characteristics of the profile selector algorithm and the ESCO ontology (Shakya & Paudel, 2019).

Tab. 3. Characteristics of selector algorithm and ESCO ontology

Profile selection algorithm	ESCO ontology
Every time that a search is done, an standardization of the profiles is carried	Classifies abilities in the US job market
Takes abilities and knowledge that appear in the job offer into account	The abilities are defined and classified in the ontology
Two search criteria	Multicriteria
It conveys 8 similarity metrics	Uses “Similarity score” metric
Tested in Spanish and English	Functional for European languages

As it can be seen on Table 3, the methodologies have different characteristics. The strong points of the proposed methodology in this document is the standardization of the elements, the use of the eight metrics, and the sorting of the evaluated profiles. It can be seen that the processing prior to the analysis of similarity metrics is very important, since, when standardizing profiles, the degree of similarity is greater, because only important information remains in the profile. At the same time, the eight metrics are taken into account for the evaluation and sorting of the profiles, this is in favor of having several experts that qualify in different ways when the evaluation is done and the sorting of the profiles that helps provide a quantitative perspective (by using the metric evaluation) on the found profiles.

5. CONCLUSIONS AND FUTURE WORK

According to our results, it can be said that it is possible to help a recruiter find the most suitable profile for a job offer, this is due to the reduction of a search range, allowing the recruiter to focus on the best evaluated profiles, besides, the search process is done in a few seconds, which means it is highly reduced, taking into consideration that a traditional recruitment method can take days or even weeks.

The disadvantage of this methodology is that it does not take into account the context of the profile. This means that a high degree similarity can be found in profiles within different areas. At the same time, this can be improved by implementing a semantic analyzer to process the elements before searching for the similarity degrees.

Future work focuses on the implementation of a stemming process. This process can be carried with the algorithm proposed by Porter (1980), the goal is to improve the similarity degree when looking for terms with a common root, this is because this kind of terms have similar meanings. Likewise, the content of the profiles can be segmented even more, which would allow a multicriteria search. At last, its functionality can be widened to other languages as long as the proper dictionary is built to be able to carry the standardization of the profiles.

REFERENCES

Baccour, L., Alimi, A., & John, R. (2014). Some notes on fuzzy similarity measures and application to classification of shapes, recognition of arabic sentences and mosaic. *IAENG International Journal of Computer Science*, *41*(2), 81–90.

Behara, K., Bhaskar, A., & Chung, E. (2018). Levenshtein distance for the structural comparison of od matrices. *40th Australasian Transport Research Forum (ATRF)*. Darwin.

Bisandu, D., Prasad, R., & Liman, M. (2018). Clustering news articles using efficient similarity measure and n-grams. *International Journal of Knowledge Engineering and Data Mining*, *5*(4), 333–348. doi:10.1504/IJKEDM.2018.095525

Cheatham, M., & Hitzler, P. (2013). String similarity metrics for ontology alignment. *International Semantic Web Conference*, *8219*, 294–309. doi: 10.1007/978-3-642-41338-419

Deng, Y., Lei, H., Li, X., & Lin, Y. (2018). An improved deep neural network model for job matching. *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 106-112. doi:10.1109/icaibd.2018.8396176

Derous, E., & Fruyt, F. D. (2016). Developments in Recruitment and Selection Research. *International Journal of Selection and Assessment*, *24*(1). doi:10.1111/ijsa.12123

Dice, L. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302. doi:10.2307/1932409

Esch, P., & Mente, M. (2018). Marketing video-enabled social media as part of your e-recruitment strategy: Stop trying to be trendy. *Journal of Retailing and Consumer Services*, *44*, 266–273. doi:10.1016/j.jretconser.2018.06.016

Esch, P., Black, J., & Ferolie, J. (2019). Marketing AI recruitment: The next phase in job application and selection. *Computers in Human Behavior*, *90*, 215-222. doi:10.1016/j.chb.2018.09.009

Gali, N., Mariescu-Istodor, R., Hostettler, D., & Fränti, P. (2019). Framework for syntactic string similarity measures. *Expert Systems with Applications*, *129*, 169–185. doi:10.1016/j.eswa.2019.03.048

González-Eras, A., & Aguilar, J. (2019). Determination of Professional Competencies Using an Alignment Algorithm of Academic Profiles and Job Advertisements Based on Competence Thesauri and Similarity Measures. *International Journal of Artificial Intelligence in Education*, *29*(4), 536–567.

Guo, X., Jerbi, H., & O'Mahony, M. (2014). An analysis framework for content-based job recommendation. In *International Conference on Case-Based Reasoning 2014*. Cork, Ireland.

Huang, A. (2008). Similarity measures for text document clustering. *New Zealand Computer Science Research Student Conference*, *6*, 49–56.

I´m Talenty (n.d.). *Intelligent platform for entailment student*. Retrieved January 10, 2019 from https://imtalenty.com/login.xhtml

Kerzendorf, W. (2019). Knowledge discovery through text-based similarity searches for astronomy literature. *Journal of Astrophysics and Astronomy*, *40*, 1–7. doi:10.1007/s12036-019-9590-5

Kessler, R., Béchet, N., Roche, M., Torres-Moreno, J., & El-Bèze, M. (2012). A hybrid approach to managing job offers and candidates. *Information Processing and Management*, *48*, 1124–1135. doi:10.1016/j.ipm.2012.03.002

Kondrak, G. (2005). N-gram similarity and distance. *String Processing and Information Retrieval*, *12*, 115–126. doi:10.1007/11575832_13

Liu, Y., Qin, K., Rao, C., & Mahamadu, M. (2017). Object-parameter approaches to predicting unknown data in an incomplete fuzzy soft set. *International Journal of Applied Mathematics and Computer Science*, *27*(1), 157–167. doi:10.1515/amcs-2017-0011

Pappis, C., & Karacapilidis, N. (1993). A comparative assessment of measures of similarity of fuzzy values. *Fuzzy Sets and Systems*, *56*(2), 171–174. doi:10.1016/0165-0114(93)90141-4

Porter, M. (1980). An algorithm for suffix stripping. *Program*, *40*, 211–218.

Sandhya, N., Lalitha, Y., Govardhan, A., & Anuradha, K. (2008). Analysis of similarity measures for text clustering. *Computer Science Journals*, *2*(4), 1–10.

Sedgewick, R., & Wayne, K. (2011). *Algorithms*, 4th Edition (pp. 244–336). Princenton.

Shakya, A., & Paudel, S. (2019). Job-Candidate Matching using ESCO Ontology. *Journal of the Institute of Engineering*, *15*(1), 1–13.

Applied Computer Science, vol. 16, no. 1, pp. 33–40
doi:10.23743/acs-2020-03

Submitted: 2020-02-26
Revised: 2020-03-10
Accepted: 2020-03-20

modeling of mechanisms, working surface meshing

Denis RATOV [0000-0003-4326-3030]*, *Vladimir LYFAR* [0000-0002-7860-9663]*

MODELING TRANSMISSION MECHANISMS WITH DETERMINATION OF EFFICIENCY

Abstract

Continuing previous studies, reviewed by modeling transmission mechanisms with the definition of one of the major criteria of transmission efficiency – the efficiency of the gearing, which depends on the geometry and kinematics of the working surfaces of the engagement and position of the contact point on the engaging surface in the object of the simulation model the designed transmission.

1. INTRODUCTION

Every modern machines and mechanisms are in their design different gear mechanisms that are parts with a complex profile. Using the full-scale simulation to automate the optimal design make it possible to transfer the process of testing actually made arrangements for testing and analysis of the simulation model, which significantly saves material and time resources for the preparation and introduction of modern machinery or equipment and guarantee their quality and reliability in the process. Using a simulation model is possible in the construction of adequate mathematical models that represent the working process of engagement and allow a comparative analysis of efficiency designed for transmission of the transmission device.

In performance gear machines usually estimated quality indicators (Gribanov, 2003) – criteria characterizing locally kinematic and hydrodynamic phenomena in the tooth contact area. Geometrical parameters of transmission loss significantly affect meshed (Gribanov, 2003; Gribanov & Ratov, 2011). The spatial gears

* Volodymyr Dahl East Ukrainian University, Faculty of Information Technology and Electronics, Department of Programming and Mathematics, Tsentralnyi Ave., 59A, Severodonetsk, Luhansk Oblast, Ukraine, 93400, denis831102@gmail.com, lyfarva61@gmail.com

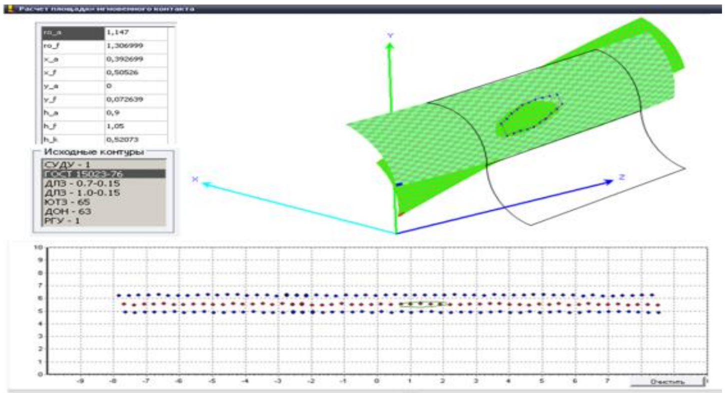
longitudinal sliding provides additional frictional losses in gearing. Therefore, an important criterion for evaluating the performance of a transmission criterion characterizing the losses in engagement. Such losses can be estimated efficiency engagement coefficient which depends on the geometry and kinematics of the working surfaces of the engagement position of the point of contact on the engagement surface.

The aim of the article is to develop a refined model for calculating efficiency spatial transmission operatively engaged with the contact interactions and side surfaces of the teeth, taking into account the rolling speed and the slip contact pads.

2. MODELING THE OBJECT OF RESEARCH

To study the efficiency of spatial transmission using a mathematical model of the process of formation of teeth on the primary hyperboloidal surfaces (Ratov & Balitska, 2007; Gribanov & Ratov, 2009). The model is based on the mutual bending of producing and manufactured surfaces. For a description of the model have been met:

1. Calculation of initial engagement circuit (Ratov & Balitska, 2007).
2. Calculation of theoretical initial spatial transmission surfaces, which are one-sheeted hyperboloids of rotation (Gribanov & Ratov, 2009; Nosko, Shyshov & Ratov, 20140.
3. Preparation of the radius-vector form of the equations describing the lateral surface of the tooth (Gribanov, 2003; *Instructional "Gleason" company materials*, 2001).
4. Model working engagement spatial wheels (Gribanov, 2003; Gribanov & Ratov, 2009). Figure 1 shows a computer implementation of working engagement simulation.



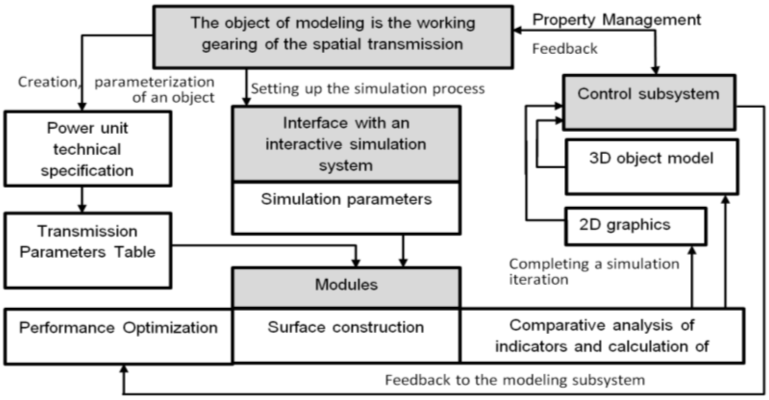
Расчет площадки мгновенного контакта

ro_a	1,147
ro_f	1,306999
x_a	0,392699
x_f	0,50526
y_a	0
y_f	0,072639
h_a	0,9
h_f	1,05
h_k	0,52073

Исходные контуры: ГОСТ 15023-76

Fig. 1. Computer implementation working engagement simulation

Based on the simulation-object modeling mathematical model obtained in the system of the hybrid parametric solid modeling was performed SolidWorks working engagement teeth contacting spatial transmission and a model of the worm gear and the gear transmission with spatial developed (Fig. 3). The structure and steps of construction in the modeling system is shown in Figure 2.



The object of modeling is the working gearing of the spatial transmission
Property Management
Feedback
Creation, parameterization of an object
Setting up the simulation process
Control subsystem
Power unit technical specification
Interface with an interactive simulation system
3D object model
Simulation parameters
Transmission Parameters Table
2D graphics
Completing a simulation iteration
Modules
Performance Optimization
Surface construction
Comparative analysis of indicators and calculation of
Feedback to the modeling subsystem

Fig. 2. Structure – modules working gears meshing spatial modeling system

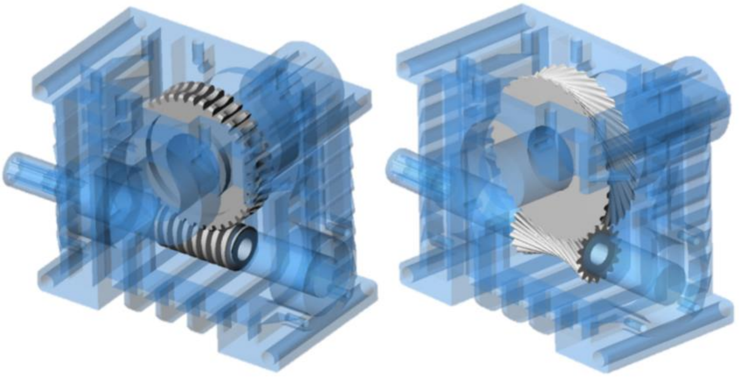


Fig. 3. Worm gearbox and gearbox with developed spatial gear

3. MODEL FOR DETERMINING EFFICIENCY COEFFICIENT

For evaluation and **comparative** analysis of the synthesized helical gear form the model for calculating the efficiency of the spatial transmission. We express this ratio as the ratio of the elementary works on the driven and the driving gear:

$$\eta = \frac{A_1}{A_2} \tag{1}$$

where: $A_1 = F_1' \cdot V^{(1)} A_2 = F_2' \cdot V^{(2)} V^{(1)} V^{(2)}$ – the elementary work on the driven and driving gears, respectively, $V^{(1)}, V^{(2)}$ – the circumferential speed of the wheels (Figure 4b), F_1', F_2' – the projection of the normal force F_n on $V^{(1)}$ and $V^{(2)}$.

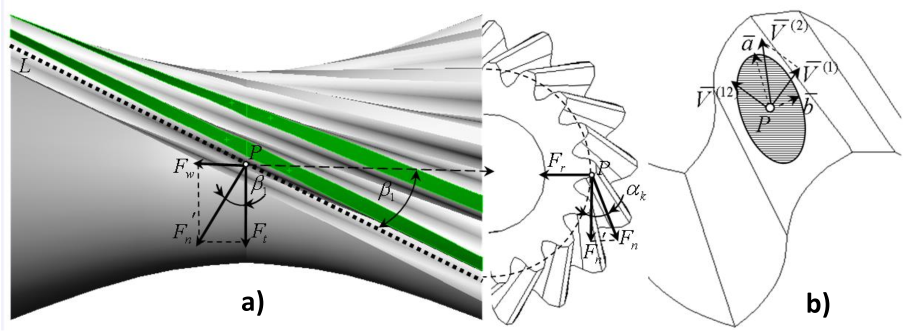


Fig. 4. a) The forces acting in the engagement F_n normal force, F_t circumference force, F_r – radial force, F_w – axial force; b) vectors peripheral speeds

Construct a mathematical model for determining the efficiency of the spatial transmission is made in consideration of contact interactions side surfaces of the teeth, taking into account the rolling speed and the slip contact pads.

We define the forces at the wheel and pinion helical gear (Fig. 4a). The normal force F_n between the surfaces of the gear teeth and pinion assumed applied to the nominal point of contact P, and its projection F_n' on a common tangent plane to the primary hyperboloidal aksoids at point P coincides with the direction normal to the teeth and is equal to:

$$F_n' = F_n \cos \alpha_k \tag{2}$$

where: α_k – in nominal pressure angle of the teeth the point of contact.

Hoop forces F_t at the wheel (all parameters with index 1) and gear (all parameters with index 2) is equal to the projection F_n' (2) in the direction perpendicular to the generator hyperboloid L (Figure 4a):

$$F_{t1} = F_n' \cos \beta_1; \; F_{t2} = F_n' \cos \beta_2 \tag{3}$$

where: β_1, β_2 – angles of generators.

Fig. 4b shows the circumferential velocity vectors point P, lying in the tangent plane. The vector of relative moving speed of the contact points of the active surfaces of the teeth (vector relative sliding velocity) will be:

$$\bar{V}^{(12)} = \bar{V}^{(2)} - \bar{V}^{(1)} \tag{4}$$

The magnitude of this vector:

$$V^{(12)} = V^{(2)} \sin \beta_2 - V^{(1)} \sin \beta_1 \tag{5}$$

We take into account the condition of equality of vectors $V^{(1)}$ and $V^{(2)}$ projections on the normal direction:

$$V^{(2)} = V^{(1)} \frac{\cos \beta_1}{\cos \beta_2} \tag{6}$$

Longitudinal sliding of the presence of the normal force between the teeth causes the frictional force directed oppositely sliding velocity and direction of the deflection force.

Consequently, the normal force on the projection plane is also tangential to change its direction, deviating at angle of friction ρ (Gribanov, 2003; Shishov, Velichko & Karpov, 2009):

$$f = \mathrm{tg}\rho \tag{7}$$

where: f – coefficient of sliding friction at the contact points the working surfaces of gear pair.

The projection of the normal force F_n on the velocity $V^{(1)}$ and $V^{(2)}$ direction are of the form:

$$F_1' = F_n \cos \alpha_k \cos(\beta_1 - \rho);\ F_2' = F_n \cos \alpha_k \cos(\beta_2 - \rho) \tag{8}$$

Substituting (6) and (8) into (1) we obtain:

$$\eta = \frac{F_1' \cdot V^{(1)}}{F_2' \cdot V^{(2)}} = \frac{F_n \cos \alpha_k \cos(\beta_1 - \rho) \cdot V^{(1)}}{F_n \cos \alpha_k \cos(\beta_2 - \rho) \cdot V^{(2)}} = \frac{\cos(\beta_1 - \rho) \cdot \cos \beta_2}{cos(\beta_2 - \rho) \cdot \cos \beta_1} \tag{9}$$

After standard identity transformations, taking into account (7), the efficiency of spatial transmission (9) takes the form:

$$\eta = \frac{(\cos \beta_1 \cdot \cos \rho + \sin \beta_1 \cdot \sin \rho) \cdot \cos \beta_2}{(\cos \beta_2 \cdot \cos \rho + \sin \beta_2 \cdot \sin \rho) \cdot \cos \beta_1} = \frac{1 + \tan \beta_1 \cdot f}{1 + \tan \beta_2 \cdot f} \tag{10}$$

Rolling bodies slidably most complete linkage geometry and the force interaction in contact bodies considered in determining the coefficient of sliding friction in (Shishov, Velichko & Karpov, 2009):

$$f = \frac{0.09 \cdot q_n^{0.1} \cdot \left(10 + \lg \frac{HB \cdot R_a}{E_r} |\chi_r|\right)}{\left(V^{(12)}\right)^{0.35} \cdot \left(V^{(\Sigma)}\right)^{0.1} \cdot \nu^{0.07}} \cdot |\chi_r|^{0.25} \tag{11}$$

where: q_n – the contact load per unit length of the line (N/cm),
HB – hardness of less solid bodies in contact (N/cm^2),
$V^{(12)}$ – normal to the contact line relative sliding velocity (cm/s),
$V^{(\Sigma)}$ – total rolling velocity (cm/s),
E_r – reduced modulus (N/cm^2),
R_a – the roughness value of the most solid (cm),
ν – oil viscosity (cSt),
χ_r – reduced curvature.

Load per unit length of the line of contact can be found from the relationship:

$$q_n = \frac{T}{r \cdot L \cdot \varepsilon} = \frac{T \cdot \cos\beta_1}{r \cdot \varepsilon \cdot b_w} \tag{12}$$

where: T – torque on the input shaft,
r – the distance from the contact point to the axial line of the shaft,
L – length of the line of contact,
ε – the overlap coefficient.

In view of formula (12) and the resistance to scoring coefficient represented in (Gribanov, 2003; Gribanov & Ratov, 2009; Ratov, & Lyfar, 2019), where it is expressed as a ratio of the relative sliding velocity teeth surfaces to the total rolling rate of contact points:

$$K_v = \frac{V^{(12)}}{V^{(\Sigma)}} \tag{13}$$

coefficient of sliding friction (11) finally becomes

$$f = 0.09 \cdot \left(\frac{T \cdot \cos\beta_1}{r \cdot \varepsilon \cdot b_w}\right)^{0.1} \cdot |\chi_r|^{0.25} \cdot \frac{\left(10 + \lg\frac{HB \cdot R_a}{E_r} \cdot |\chi_r|\right)}{(K_v)^{0.35} \cdot \left(V^{(\Sigma)}\right)^{0.45} \cdot \nu^{0.07}} \tag{14}$$

The turns of the worm in the worm gear slip when driving on the wheel teeth. Large slip is the cause of the wear and seizing of such transfers, reducing their efficiency. Model for calculating efficiency worm gear similar construct, the calculation of the efficiency screw pair, since the friction conditions they are identical (Shishov, Velichko & Karpov, 2009):

$$\eta = 0.95 \frac{\mathrm{tg}\gamma}{\mathrm{tg}(\gamma + \rho)} \tag{15}$$

where: 0.95 – factor takes into account the energy loss in mixing oil for lubrication dipping,
γ – lifting pitch helix angle ($5° - 20°$),
ρ – the friction angle.

Considering equation (7), the expression for calculating the efficiency a worm gear (15) takes the form:

$$\eta = 0.95\,\frac{\mathrm{tg}\gamma\cdot(1-\mathrm{tg}\gamma\cdot f)}{\mathrm{tg}\gamma+f} \qquad (16)$$

4. THE METHODOLOGY AND THE RESULTS OF RESEARCH

Using the equations of the surfaces in the radius-vector form $\bar{r}\,(\nu,\varphi)$ (Gribanov, 2003; *Instructional "Gleason" company materials*, 2001), the equation for K_ν (Gribanov, 2003; Gribanov & Ratov, 2009; Ratov, & Lyfar, 2019), $V^{(\Sigma)}$ (Gribanov & Ratov, 2009), which are defined using the coefficients of the quadratic forms of the surface *E*, *F*, *G*, equation given curvature χ_r, the resultant sliding friction coefficient (14), taking into account the efficiency of the equation spatial transmission (10) and worm gear (16) in Computer Algebra System Wolfram Mathematica 8 construct a mathematical model for calculating efficiency spatial transmission operatively engaged (at the angle of engagement) and worm gear (Fig. 5). In this case T = 23000 Ncm; HB =2800 N/cm^2; E_r = 203.9·10^3 N/cm^2; R_a= 4·10^{-5} cm; ν = 20 cSt.

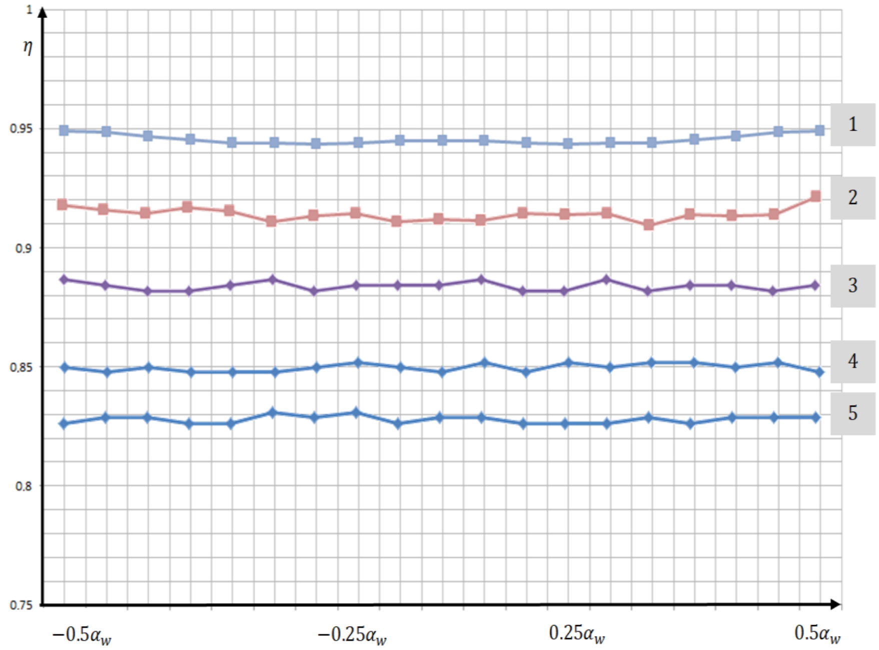


Fig. 5. The efficiency of the test gear

Fig. 5 is a graph 1, 2 – efficiency spatial transmission schedule 3–5 – efficiency worm gear. The average efficiency of the theoretical value spatial transmission №1 – $\eta_{\text{cp}} = 0.945$, spatial transmission №2 – $\eta_{\text{cp}} = 0.9114$, worm gear №3 – $\eta_{\text{cp}} = 0.8845$, worm gear №4 – $\eta_{\text{cp}} = 0.8538$, №5 worm gear $\eta_{\text{cp}} = 0.8334$.

Efficiency worm gear increases with the number of turns of the worm (increase divider angle γ). In the worm gear №5 number of turns is 1, y transmission №3 – 4. Increasing the number of turns γ decreases the rigidity of the divider angle of the worm, and increasing the number of teeth increases the transmission distance between the supports. Increase efficiency spatial transmission №1, № 2 caused a decrease in friction coefficient f of the contact surfaces (friction angle decreases ρ), which depends on the reduction ratio K_v (reduction of relative sliding velocity $V^{(12)}$ and the total velocity $V^{(\Sigma)}$ increase lateral rolling surfaces).

5. CONCLUSIONS

1. Developed a refined model for calculating efficiency engagement with the contact interactions and side surfaces of the teeth, taking into account the rolling speed and the slip contact pads.
2. Conducted implementation of simulation and object modeling. The results of computing experiment, which showed an increase of the value of efficiency gireboloidnoy investigated gearbox transmission.

REFERENCES

Gribanov, V. (2003). *Theory of hyperboloid gears* (p. 272). Lugansk, Ukraine: Publishing House of the East Ukrainian National University named after V. Dahl.

Gribanov, V., & Ratov, D. (2009). Simulation of hyperboloid gears modelling. *TEKA, Commission of Motorization and Energetics in Agriculture: Polish Academy of sciences*, *9*(1), 54–60.

Gribanov, V., & Ratov, D. (2011). On the coefficient density active abutment surfaces of teeth of helical gear. *Journal of East Ukrainian National University named after Vladimir Dahl*, *11*(165)**,** 7–15.

Instructional "Gleason" company materials. (2001). Design of bevel and hypoid gears. USA.

Nosko, P., Shyshov, V., & Ratov, D. (2014). Helical gear train load capacity criterion. *TEKA, Commission of Motorization and Energetics in Agriculture: Polish Academy of sciences*, *14*(1), 182–190.

Ratov, D., & Balitska, T. (2007). Numerical multicriteria synthesis of Novikov DLZ gears. *Journal of East Ukrainian National University named after Vladimir Dahl*, *12*(118), 52–56.

Ratov, D., & Lyfar, V. (2019). Determination of the density factor of the fit in mathematical modeling of the working gear of spatial transmissions. *Mathematical modeling in economics. Section mathematical and information models in economics*, *4*, 50–60.

Shishov, V., Velichko, N., & Karpov, A. (2009). *Highload helical gears* (p. 240). Lugansk: Publ EUNU. Dal.

Applied Computer Science, vol. 16, no. 1, pp. 41–59
doi:10.23743/acs-2020-04

Submitted: 2020-02-01
Revised: 2020-02-19
Accepted: 2020-03-13

Deep Learning, topic modeling, Text Mining, ADR, NMF

Behnaz ESLAMI[*], *Mehdi HABIBZADEH MOTLAGH*[**],
Zahra REZAEI[***], *Mohammad ESLAMI*[****],
Mohammad AMIN AMINI[*****]

UNSUPERVISED DYNAMIC TOPIC MODEL FOR EXTRACTING ADVERSE DRUG REACTION FROM HEALTH FORUMS

Abstract

The relationship between drug and its side effects has been outlined in two websites: Sider and WebMD. The aim of this study was to find the association between drug and its side effects. We compared the reports of typical users of a web site called: "Ask a patient" website with reported drug side effects in reference sites such as Sider and WebMD. In addition, the typical users' comments on highly-commented drugs (Neurotic drugs, Anti-Pregnancy drugs and Gastrointestinal drugs) were analyzed, using deep learning method. To this end, typical users' comments on drugs' side effects, during last decades, were collected from the website "Ask a patient". Then, the data on drugs were classified based on deep learning model (HAN) and the drugs' side effect. And the main topics of side effects for each group of drugs were identified and reported, through Sider and WebMD websites. Our model demonstrates its ability to accurately describe and label side effects in a temporal text corpus by a deep learning classifier which is shown to be an effective method to precisely discover the association between drugs and their side effects. Moreover, this model has the capability to immediately locate information in reference sites to recognize the side effect of new drugs, applicable for drug companies. This study suggests that the sensitivity of internet users and the diverse scientific findings are for the benefit of distinct detection of adverse effects of drugs, and deep learning would facilitate it.

[*] Islamic Azad University, Science and Research Branch, Department of Computer Engineering, Islamic Azad University, Tehran, Iran, behnazeslami30@gmail.com
[**] P/S/L Group, 1801 McGill College Ave, Montreal, Quebec H3A 2N4, Montreal, Canada
[***] University of Kashan, Department of Computer and Electrical Engineering, Isfahan Province, Qotb-e Ravandi Blvd, Kashan, Iran
[****] Islamic Azad University of Qazvin,Department of Computer Engineering, Qazvin, Iran
[*****] Islamic Azad University of Jasb, Department of Computer Engineering, Markazi, Iran

1. INTRODUCTION

The Adverse Drug Reaction (ADR) is defined as "an undesirable effect". The 'side effect' does not have the exact terminology for inadvertent and secondary effect, observed during therapy. In fact, the interpretation of term "side effect" may vary between two different individuals. However, adverse drug reactions could be considered as the result of toxicity from all kinds of drugs. Apparently, 3 to 7% of all hospitalizations have been due to adverse drug reactions (Kongkaew, Noyce & Ashcroft, 2008). And ADRs noticeably increase patient's hospitality costs (Sultana, Cutroneo & Trifirò, 2013; Miranda, 2018). According to the annual report of the Agency for Healthcare Research and Quality, over 770,000 patients were injured and/or died in hospitals due to adverse drug reactions in each year (Rison, 2013).

Based on similar singling pathways and cellular structures, involved in normal or abnormal conditions, the same expectation on side effect and actual treatment effect would probably make the uniform pattern for medication. The goal of any drug administration needs to focus on differentiation between negative and positive effect of targeted drug as much as possible, which is required to be tested case by case. The focus of our study is to investigate into appropriate dosage of drugs, since the biological response of each individual to different medication may be various, i.e. one specific drug probably has unexpected destructive effect on one individual, while it is safe for others, thus the interaction between drug and cells need to be adjusted, whose index is normalization of drug dosages per case. Fortunately, there have been available reports for drug interaction in social media which help public have good understanding of side effect. For instance, it has been reported that aspirin and warfarin interfere with clot formation in blood vessels and the subsequently bleeding time would take longer. Another example is the feedback of food or herbs to drugs which modifies their effects, i.e. it has been reported that the level of cholesterol in the circulatory system is reduced by statins however, high fat diets have an opposite effect on blood cholesterol level. Also, St. John's Wort could make bipolar individual hyperactive in spite of consumption of the antidepressant drug (Bordet, Gautier, Louet, Dupuis & Caron, 2001).

It takes a well-trained reader a lot of time to screen ADRs by looking through relevant literatures without using a machine reader. Therefore, it is crucially valuable for experts to benefit from automated system in order to find ADRs in publications as fast and efficiently as possible (Classen, Pestotnik, Evans, Lloyd & Burke, 1997). The detection of ADRs have not been initially well-structured and just obtained through communication between health professionals and patients or published case reports, available in MEDLINE, PubMed or other publicly available datasets (Rison, 2013; Vallano et al., 2005). Hence, society needs an alternative approach to detect side effects of the clinical medications. The social media is capable of producing novel and reliable data sources for the side effects of drugs.

In fact, through the social media, special events in the field of health could be identified and managed. “Ask a patient” is the web page that allows patients to share and compare medication experiences, and was granted Webby Award for the best website in the Pharmaceutical Category in 2012. The “Ask a patient” database contains more than 4,000 chemically prepared and prescribed drugs, approved by FDA’s Center for Drug Evaluation and Research.

Comments over prescription or the counter drugs, found in this web page, would be based on fine-tuned search criteria (age, gender, symptom, etc.). However, the difference between written and oral language in social media creates some noises. Also, lack of a suitable structure and imbalance data in drug groups are considered as important challenges in classification of data, retrieved from social media. Accordingly, in spite of richness of health-related data in social media, it seems not to be practical to use this type of data for the purpose of ADR detection.

In this study, we identify drug side effect based on three main criteria:

1. An automated deep learning was applied to extract features from social media. The comments of “Ask a patient” website’s users, were processed to describe side effects and thus reduce the difference between written and oral language and dampen down the noise effect.
2. The efficacy of deep learning method in classification of data from “Ask a patient” was approved by the quality of the outcome. The results showed that deep learning performance benefits from high accuracy and speed, simultaneously.
3. Advantage and disadvantage of each comment were compared with those of already reported ones in Sider and WebMD web pages. In order to achieve that, deep learning method HAN (Yang et al., 2016) was employed to classify users’ comments. Then, the non-monitoring method (NMF) of topic modeling was administered to determine specific topics in each group of drugs.

2. RELATED WORKS

Some studies have hitherto investigated into the side effect of drugs using social media as tool. For example, Sarker and Gonzalez highlighted the importance of combined usage of advanced NLP-based information generation and traditional text classification (Support Vector Machine, Naïve Bayes and Maximum Entropy) to accurately detect and classify sentences concerning ADR (Sarker & Gonzalez, 2015). Aligned with that, Ho et al. suggested the automated detection of data related to ADR by searching relevant database; they prepared a systematic review and concise information about suitable approach to envisage ADEs, pointed out in social media (Ho, Le, Thai & Taewijit, 2016).

Also, Ginn and coworkers applied two supervised machine learning approaches (NB and SVM) on a wide range of annotated medications in association with ADR tweets (Ginn et al., 2014). Although, the classifier showed moderate performance, it was considered as the base for future development in advanced techniques. Aligned with this approach, they used Convolutional Neural Networks (CNN) model, which applied word2vec embedding for classification of Twitter comments. In contrast to other models, their proposed model not only used a small fraction of features for data collection, but also showed high performance in text classification procedures (Akhtyamova, Alexandrov & Cardiff, 2017a). Recent attempts have been made to benefit from specific type of deep learning to enhance quality of ADR discovering through extraction of sentences and entities, available in social media. Gupta et al. suggested a two-step method to extract pointed out adverse event, i.e. it initially predicts drug with regard to input contexts, unsupervisedly, and then it repeats same direction in a supervised way (Gupta, Pawar, Ramrakhiyani, Palshikar & Varma, 2018). In parallel, Tan et al. offered the summary of data base and automated systems to support ADRs detection (Tan et al., 2016). Also, Harpaz et al. presented the synopsis on using text mining for the purpose of Adverse Drug Events (ADEs) detection, in publicly available literature or web pages (Harpaz et al., 2014).

In addition, Lee and colleagues put forward a semi-supervised CNN-based framework to classify the adverse drug event (ADE) in Twitter. A Twitter dataset was used in PSB 2016 Social Media Shared Task, leading to high performance classification of ADE with 9.9% F1-Score (Lee et al., 2017). It is good to be pointed out that ADE detection surveillance systems require small number of labeled instances. Also, Akhtyamova et al, presented a CNN-based architecture, composed of numerous parameters to predict adverse drug reaction based on the quantity of votes (Akhtyamova, Alexandrov & Cardiff, 2017b). They utilized a large scale of medical dataset, derived from medical websites, in order to evaluate the mode of performance. In contrast to previously reported networks, the proposed end-to-end model does not require handcrafted features and data pre-processing, and it resulted in an enormous improvement in standard CNN based methods.

Finally Rezaei et al, suggested three methods for preprocessing of data analyses and used numerous deep learning methods for text classification. Compared to current deep learning-based networks, their results showed that the FastText, CNN, and HAN were much faster and more accurate. According to deep learning models, they suggested the approach of end-to-end, in which artificial attribute and preprocessed information are not necessary. The obtained results demonstrated that the proposed models would significantly improve the performance of baseline methods for different datasets. They noticed that increasing batch size during training steps considerably reduced the learning rate in the network. Conversely, they tested various optimizers including SGD, RMS, and Adam in their custom datasets, and found that Adam shows better results compared to RMS and SGD (Rezaei, Ebrahimpour-Komleh, Eslami, Chavoshinejad & Totonchi, 2020).

This study aims to investigate the written topic modeling of typical users and identify the changes in comments, which have been reported from 10 years ago. We designed a model that provides researchers with immediate capability of analyzing comments through combined deep learning methods.

3. METHOD

This paper is organized into two sections; classification and extraction of topics (Fig. 1).

3.1. Classification

3.1.1. Data Sources

Prior to data collection, we selected a set of interesting drugs, which were likely to have a large number of associated comments in “Ask a patient” database. We chose drugs that were prescribed for chronic diseases and syndromes, i.e. the medication with high prevalent prescription and referred comments. The names of the medications were reported in separate classes (Anti-depressant drugs, Anti-Pregnancy drugs and Gastrointestinal drugs) in figures 2 to 4.

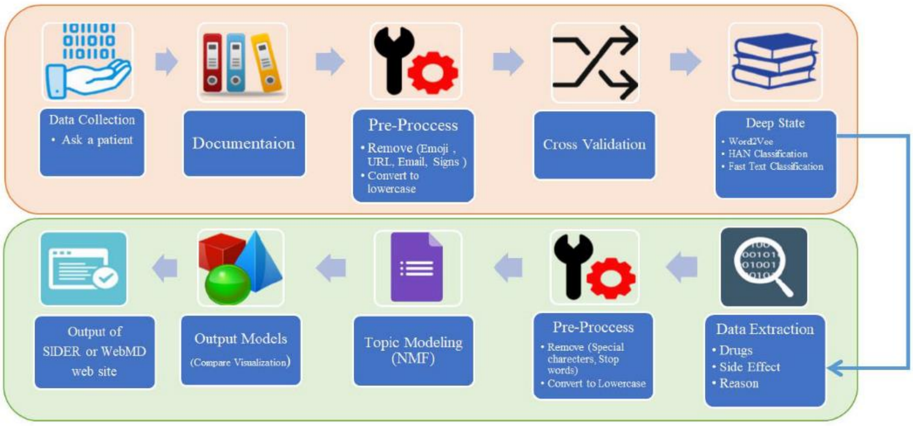


Fig. 1. The workflow of the proposed deep learning based strategy is illustrated

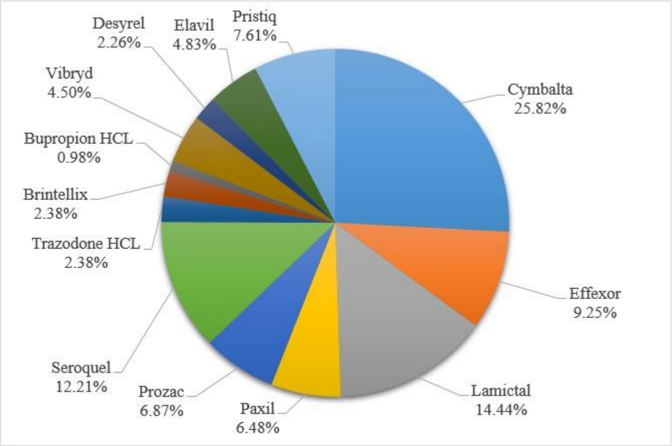


Fig. 2. Anti-Depressant Medicines Side effects (4929 Comments)

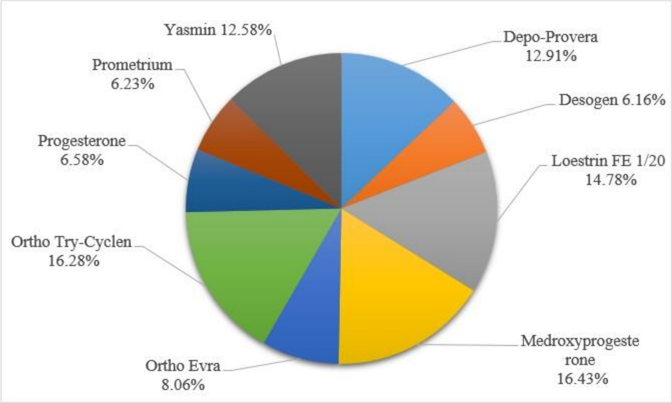


Fig. 3. Anti-Pregnancy Medicines Side effects (4149 Comments)

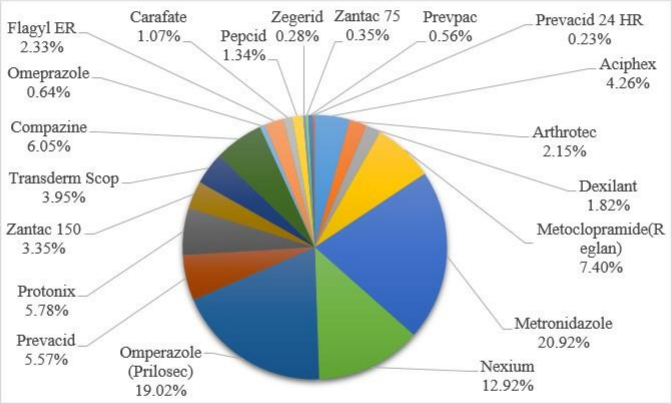


Fig. 4. Digestion Medicines Side effects (3995 Comments)

3.1.2. Preprocessing

The pre-processing comments in both data are done as follows:

- Data shuffling,
- Converting all uppercase words to lowercase ones,
- Elimination of special characters like: @, !, /, *, $ and etc.,
- Removal of stop word: at, of, the, … ,
- Correction of words with repeated characters like: pleaseeeeeeeeee and/or yessss,
- Conversion of contractions to base format like: I'm → I **am**,
- Lemmatization: I started taking almost two months ago. → I **start take** almost two months ago.

3.1.3 Cross Validation

In order to achieve the best performance with regard to new data, we wished to find the appropriate values of the complexity parameters, leading to optimal model. If the amount of data was high, the procedure would have been divided into three subsets; the training, the validation and the test sets. Among the diverse complex models that have been trained, we selected the one that had the best predictive and effective performance, and was confirmed by the data in the validation set. However, the data supply was limited for training and test set, which led to the increase of the generalized error. Thus, cross validation was applied to reduce these types of error and prevent over-fitting. The data distribution for each group is shown in Table 1.

Tab. 1. Distribution of data in Cross-Validation phase

Medicines Category	Train Phase Docs	Test Phase Docs	Validation Phase Docs
Neurotic and Anti Depression Medicines	4437	492	982
Anti-pregnancy Medicines	3735	414	828
Digestion Medicines	3596	399	798

3.1.4. Deep Classification

The applied methods for data classification are HNN (Yang et al., 2016) and FastText (Joulin, Grave, Bojanowski & Mikolov, 2016) with similar word2vec section. Once word2vec generated, this file would be used for further investigations.

3.1.4.1. HAN Method

Hierarchical Attention Network (HAN) has two distinctive characteristics: (I) a hierarchical structure and documents, (II) two-phase mechanism of attention, which enables HAN to differentially put words or sentences next to each other within the structure of the document. In addition to these two characteristics, HAN network is composed of quite a few parts including, i.e. a word sequence encoder, a word-level attention layer, a sentence encoder and a sentence-level attention layer. HAN works based on a positive role of sentences and document structure in modeling.

3.1.4.2. FastText Method

This method demonstrates a simple and efficient approach for classification of the texts and its expressions. Large numbers of studies show that the classification of texts with this method is faster in comparison with deep learning methods, with regard to accuracy and applied commands for training and evaluation.

Tab. 2. (HAN and FastText) Training Phase Configuration

Training Phase
Initializations:
Configuration of Distributed Parameters {Device: {**NVIDIA GEFORCE GTX 1050, RAM 16G**}}
Configuration of Optimization {Name of optimization: {“**Adam, SGD** and **RMS prob**”}}
Configuration of Loss {Name of loss-function: {“**Sigmoid**”}}
Initials {Pad_Seq_Len: {**150**},
Embedding_Dim: {**100**}, // for creating Word2Vec model
Batch_Size: {**32**, **64** and **128**},
Epochs: {**100**}},
Learning Rate: {0.1, 0.01, 0.001}
Configuration of Data Set {Datasets: {**Train.json**}}
Main ():
Select the Dataset // Based of Application and select Train part
Select the Network // A function that applies the model to a batch of documents
Create a dataset provider that loads data from the dataset // Return [**Content, Label**]
Create Training Operations
Run the Training

In terms of structure, there are two major and influential differences, as follow:

- Softmax: It is a hierarchy, based on the Huffman encoded tree structure that reduces Time Complexity O(Kd) to O(d log k), where K is number of targets and D is dimension of the hidden layer.
- N-gram features: While Bag of words is invariant to word order; it is very expensive to take simplicity into consideration. Instead, we used bag of n-gram as an additional feature to capture some partial information about local word order, which seems to be more efficient in practice (Table 2).

3.1.4.3. Evaluation Metrics

- Precision (positive predictive value) and recall (sensitivity): These metrics are appropriate fraction of retrieved samples from all and relevant instances. Application of these metrics depends on understanding and measuring of relevance.
- Accuracy: This criterion is the accuracy of the x-group classification against all items where the x-tag for investigating records is suggested by means of classification. This criterion indicates how much reliable is the classification output is reliable.
- F-measure: This criterion is a combination of call metrics and accuracy and it is used to find if it is impossible to consider special importance to each of the two criteria.
- Kappa: This criterion is often used to test the reliability of the viewer and to compare the accuracy of the system in terms of how much generated output is coincident.

Tab. 3. Evaluation metrics formula

Metrics
$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$
$Recall = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$
$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+PF+PN}}$
$\text{F-}Score = \frac{\mathrm{Precision*Recall*2}}{\mathrm{Precision+Recall}}$
$Kappa = \frac{\mathrm{Pr(a)-Pr(e)}}{\mathrm{1-Pr(e)}}$

3.2. Extracted Topics

3.2.1. Data Sources

Three classes of drugs have been consumed between 2008 and 2018 in figures 2 to 4.

3.2.2. Topic Modeling

As a linear algebraic model, Non-negative Matrix Factorization (NMF) includes high-dimensional vectors and low-dimensional image. Vectors are non-negative in NMF like Principal Component Analysis (PCA). Skewing the vectors towards lower-dimensional form in NMF makes the coefficients non-negative.

The two matrices of **W** and **H**, would be obtained through original matrix A, in which **A = WH**. Also, NMF has an inborn clustering property. **A, W** and **H** represent the following information:

- **A (Document-Word Matrix):** input that shows which words appear in which documents.
- **W (Basis Vectors):** the topics (clusters) are elicited from the documents.
- **H (Coefficient Matrix):** the membership weights for the topics in each document.
- **W** and **H** are calculated by optimization of an objective function (like the *EM algorithm*), and updating both **W** and **H**, iteratively, until they are converged (Table 4).

Tab. 4. NMF topic modeling configuration

NMF Topic Modeling
Initializations:
Number of Topics: {10}
Number of Top Words: {20}
Configuration of feature extraction by using TfidfVectorizer: {
Initials: {
ngram_range: {(2, 2)},
Minimum Document Frequency (min_df): {2},
Configuration of NMF Topic Modeling Parameters and fit by TfidfVectorizer: {
components: {Number of Topics},
init: {'Scikit-Learn implementation of NMF (**including NNDSVD initialization)'},** // better for sparseness **}}}**
Run to extracting Topics

4. RESULT

4.1. Usage Model

In this study, we benefited from user's comments in "Ask a patient" to extract side effects of drugs. In general, the scale of curser that moves over texts in both FastText and HAN methods is called *Pad_Seq_Len* and we considered quantity equal to 150 for that; because, the maximum size of comments is 150 to pay more attention to the length of sentences and semantic conjugation. Moreover, the value of Embedding dim was 100. We evaluated several optimizations such as *Stochastic Gradient Descent*, *RMS probe* and *Adam*. That *Adam* shows better results (Table 5).

The value of ngram_range was chosen based on the side effects, extracted from Sider or WebMD websites. Other values such as (1, 2), (2, 3) and (3, 3) were determined but (2, 2) was the best choice (Table 6).

Tab. 5. HAN hyper parameters

Pad_Seq_Len	150
Embedding_Dim	100
Drop_Out_Prob	0.5
Loss	Sigmoid
Optimization	Adam

Tab. 6. Evaluation metrics formula

ngram_range	min_df
(2, 2)	2

4.1. Implementation Model in 3.1

In this research the used hardware includes: NVIDIA GEFORCE GTX 1050 and CPU Intel Core i7. Two methods of classification were applied against three different data groups listed in the following tables (Table 7 and 8). As shown in these tables, the best result in each method, the learning rate as well as batch size was evaluated. Also, different criteria have been tested for each type of model according to the type of data, which have been obtained in various values. For example, applying HAN method including Batch size of 128 and learning rate of 0.001 on "Ask a patient" dataset and resulting in highest accuracy (0.924) which is highlighted in Table7.

Tab. 7. Output of deep learning classification (HAN Method) on dataset

Dataset	Method	Batch Size	Learning Rate	Accuracy	Kappa	Recall	Precision	F1 Score
Ask a Patient	HAN	32	0.1	0.881	0.821	0.878	0.887	0.881
				0.883	0.842	0.881	0.885	0.882
				0.908	0.862	0.906	0.911	0.907
		64	0.01	0.889	0.833	0.887	0.891	0.888
				0.873	0.808	0.870	0.876	0.872
				0.921	0.881	0.919	0.924	0.921
		128	**0.001**	0.888	0.831	0.885	0.891	0.887
				0.879	0.818	0.879	0.878	0.879
				0.924	**0.885**	**0.921**	**0.926**	**0.923**

Tab. 8. Output of deep learning classification (FastText Method) on dataset

Dataset	Method	Batch Size	Learning Rate	Accuracy	Kappa	Recall	Precision	F1 Score
Ask a Patient	FastText	32	0.1	0.892	0.837	0.888	0.897	0.892
				0.872	0.806	0.866	0.887	0.870
				0.891	0.836	0.888	0.895	0.891
		64	0.01	0.896	0.843	0.894	0.897	0.895
				0.885	0.827	0.884	0.886	0.885
				0.899	0.848	0.898	0.899	0.898
		128	**0.001**	0.876	0.814	0.876	0.876	0.875
				0.895	0.841	0.892	0.896	0.894
				0.909	**0.863**	**0.908**	**0.909**	**0.909**

4.2. Implementation model in 3.2

Considering the output of the previous phase, the three features i.e. Side effects, reason and drug were used. Accordingly, in each class of drugs (neurotic medicines, anti-pregnancy and gastrointestinal), 10 topics with high priority were selected. As shown in tables 9 to 11, topics of each class are verbally similar.

Tab. 9. Anti-depressant Medicines Topic Modeling ("Ask a patient")

Topic #0:	Topic #1:	Topic #2:	Topic #3:	Topic #4:	Topic #5:	Topic #6:	Topic #7:	Topic #8:	Topic #9:
weight gain	dry mouth	memory loss	vivid dream	hair loss	loss appetite	brain zap	weight loss	panic attack	miss dose
extreme weight	mouth constipation	severe memory	night vivid	loss weight	nausea loss	loss libido	appetite weight	suicidal thought	dose hour
increase appetite	mouth weight	loss confusion	dream nightmare	loss memory	loss libido	zap dizziness	decrease appetite	mood swing	dose dizzy
major weight	blur vision	loss trouble	insomnia vivid	blur vision	appetite weight	dizziness brain	slight weight	anxiety panic	withdrawal symptom
massive weight	gain dry	loss weight	dream night	loss hair	mouth loss	horrible brain	loss loss	increase anxiety	dizziness miss
gain loss	nausea dry	loss loss	decrease libido	joint pain	insomnia loss	depression anxiety	nausea weight	restless leg	hour miss
constipation weight	mouth headache	loss memory	gain vivid	gain hair	headache loss	inability orgasm	week weight	weird dream	nausea dizziness
gain increase	headache dry	blur vision	dream increase	loss insomnia	taste mouth	sleep paralysis	loss weight	depression suicidal	zap miss
gain constipation	mouth sleepiness	long memory	increase dose	memory problem	increase depression	withdrawal symptom	gain weight	extreme fatigue	dose miss
rapid weight	mouth loss	gain memory	dream decrease	muscle ache	dizziness loss	zap dose	loss severe	severe panic	24 hour
gain fatigue	constipation dry	confusion memory	acid reflux	itchy scalp	trouble sleep	flu symptom	insomnia weight	lack emotion	dose day
lose weight	extreme dry	dizziness memory	dose vivid	week stop	upset stomach	zap miss	loss increase	anti depressant	headache nausea
loss libido	mouth week	slight memory	extremely vivid	extreme weight	appetite day	horrible withdrawal	loss decrease	start medication	dose vivid
gain weight	mouth dizziness	loss libido	sleep vivid	memory impairment	loss sex	zap severe	loss month	leg syndrome	dose brain
fatigue weight	mouth insomnia	loss hair	heart palpitation	dry skin	nausea vomit	dose miss	brain fog	night terror	dose night
mouth weight	mouth night	brain fog	lose weight	loss dry	day nausea	zap nausea	headache weight	trouble sleep	gain weight
slight weight	blood pressure	lack concentration	day night	make sense	fatigue loss	gain brain	hour sleep	anxiety depression	depression anxiety
gain month	appetite dry	night sleep	sleep day	vivid nightmare	increase anxiety	nausea brain	loss nausea	increase suicidal	pin needle
gain dry	sleep dry	slur speech	night loss	constipation fatigue	appetite loss	nausea constipation	loss sleep	heart race	severe withdrawal
loss sex	ring ear	mood swing	day sleep	nausea dizziness	stomach pain	extreme dizziness	delay ejaculation	anxiety increase	electric shock

Tab. 10. Anti-depressant Medicines Topic Modeling ("Ask a patient")

Topic #0:	Topic #1:	Topic #2:	Topic #3:	Topic #4:	Topic #5:	Topic #6:	Topic #7:	Topic #8:	Topic #9:
weight gain	mood swing	breast tenderness	hot flash	hair loss	birth control	panic attack	loss sex	sore breast	weight loss
gain depression	swing depression	nausea breast	flash night	loss weight	gain weight	depression anxiety	gain loss	abdominal pain	clear skin
slight weight	severe mood	extreme breast	day hot	loss appetite	blood clot	anxiety panic	swing loss	gain sore	light period
gain mood	extreme mood	slight breast	night hot	gain hair	tri cyclen	severe anxiety	fatigue loss	breast acne	period weight
swing weight	depression mood	swing breast	swing hot	anxiety depression	lose weight	severe depression	vaginal dryness	breast nausea	loss period
depression weight	swing weight	tenderness weight	low pain	depression hair	ortho tri	attack depression	sex depression	lose weight	loss loss
bloat weight	gain mood	tenderness mood	vivid dream	dry eye	control pill	attack anxiety	anxiety loss	zero sex	loss acne
gain acne	bad mood	increase appetite	light head	extreme hair	recommend birth	heart palpitation	total loss	cramp mood	increase sex
yeast infection	headache mood	severe breast	depo shot	joint pain	ortho evra	depression panic	moodiness loss	extreme fatigue	loss fatigue
extreme weight	horrible mood	tenderness nausea	depo provera	vaginal dryness	period month	suicidal thought	sex fatigue	vaginal dryness	yeast infection
gain anxiety	vaginal dryness	tenderness headache	long period	swing hair	month stop	anxiety depression	depression loss	chest pain	regular period
gain sex	swing anxiety	tenderness depression	pill day	heart palpitation	stop period	extreme anxiety	sex weight	month period	lot weight
decrease sex	major mood	tenderness increase	severe cramp	loss loss	month period	severe panic	sex mood	breast nipple	loss appetite
headache weight	swing headache	headache breast	headache nausea	heavy period	sick stomach	chest pain	sex vaginal	dry mouth	skin weight
increase appetite	anxiety mood	cramp breast	race heart	sex hair	start pill	swing depression	dryness loss	start period	period cramp
gain increase	increase appetite	tenderness loss	trouble sleep	loss acne	period heavy	extreme depression	extreme fatigue	vivid dream	decrease appetite
gain weight	swing irritability	light period	heart attack	loss extreme	make gain	anxiety weight	headache loss	fluid retention	severe depression
low sex	fatigue mood	tenderness swell	gain bloat	loss depression	heavy period	brain fog	loss loss	breast cramp	vaginal dryness
gain loss	nausea mood	miss period	fatigue mood	severe depression	blood thinner	swing anxiety	race heart	breast mood	appetite weight
gain moodiness	swing sex	gain breast	anxiety insomnia	painful intercourse	body use	headache anxiety	painful intercourse	day provera	loss libido

Tab. 11. Anti-depressant Medicines Topic Modeling (“Ask a patient”)

Topic #0:	Topic #1:	Topic #2:	Topic #3:	Topic #4:	Topic #5:	Topic #6:	Topic #7:	Topic #8:	Topic #9:
taste mouth	panic attack	dry mouth	stomach pain	joint pain	heart palpitation	chest pain	anxiety depression	blur vision	stomach cramp
metallic taste	anxiety panic	extreme dry	severe stomach	muscle pain	anxiety heart	blood pressure	severe anxiety	dizziness blu	severe stomach
dark urine	extreme anxiety	mouth headache	pain nausea	pain muscle	shortness breath	shortness breath	loss appetite	mouth blur	cramp pain
bad taste	depression anxiety	extremely dry	pain stomach	weight gain	blood pressure	pain chest	shortness breath	pain blur	cramp nausea
loss appetite	severe panic	mouth bad	bad stomach	pain joint	palpitation anxiety	pain anxiety	mood swing	vision blur	cramp diarrhea
metal taste	race heart	severe dry	pain cramp	muscle joint	hair loss	anxiety chest	depression fatigue	weight gain	nausea stomach
horrible taste	crawl skin	blurry vision	pain constipation	severe joint	brain fog	pain heart	extreme anxiety	fatigue blur	nausea vomit
nasty taste	suicidal thought	headache dry	pain bad	muscle weakness	palpitation dizziness	high blood	weight loss	sensitivity light	headache stomach
mood swing	attack anxiety	patch day	bloat stomach	brain fog	high blood	heart attack	nausea loss	extremely dry	diarrhea stomach
loose stool	think die	bad taste	headache stomach	pain pain	tightness chest	hand foot	depression loss	poor concentration	loose stool
flu symptom	severe anxiety	mouth dry	sore throat	pain severe	muscle twitch	weight gain	muscle spasm	sore throat	brain fog
horrible metallic	brain fog	dizziness dry	mouth stomach	severe headache	dizziness heart	palpitation chest	brain fog	vision anxiety	muscle cramp
light head	attack depression	mouth blur	pain bloat	body ache	headache heart	muscle pain	suicidal thought	fog blur	cramp bloat
bitter taste	heart race	mouth loss	pain anxiety	ring ear	lump throat	pain tightness	depression panic	remove patch	dark urine
upset stomach	heart rate	brain fog	pain headache	pain shoulder	anxiety attack	hair loss	sore throat	mood swing	bad stomach
mouth dark	shortness breath	light head	pain severe	leg cramp	light headedness	tightness chest	extreme fatigue	headache dizziness	cramp stomach
day day	hand foot	wear patch	diarrhea stomach	pain fatigue	trouble sleep	pain shortness	trouble sleep	extreme dry	sick stomach
extreme nausea	horrible anxiety	abdominal cramp	terrible stomach	pain swell	pain heart	race heart	ring ear	mental fog	diarrhea nausea
extreme fatigue	lose mind	muscle cramp	pain day	pain leg	light head	heart rate	major anxiety	48 hour	dizziness stomach
metalic taste	horrible panic	mouth throat	body ache	blurry vision	race heart	rapid heartbeat	race heart	weight loss	cramp severe

After extraction of these tables, all are mapped with a similar word, and meaningless topics were deleted. Figures 5, 6 and 7 show the frequency of repetition of topic models.

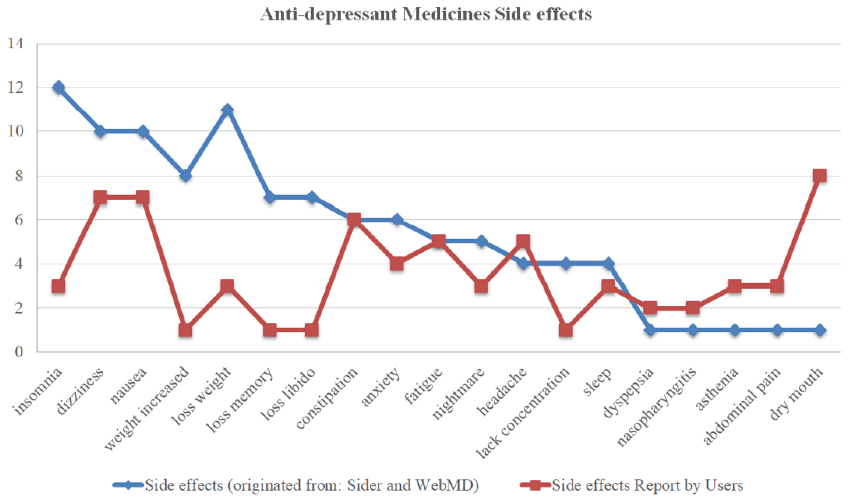


Fig. 5. Comparison of Topic Modeling of users' comments with the side effects reported on the websites of Sider and WebMD (Neurotic drugs)

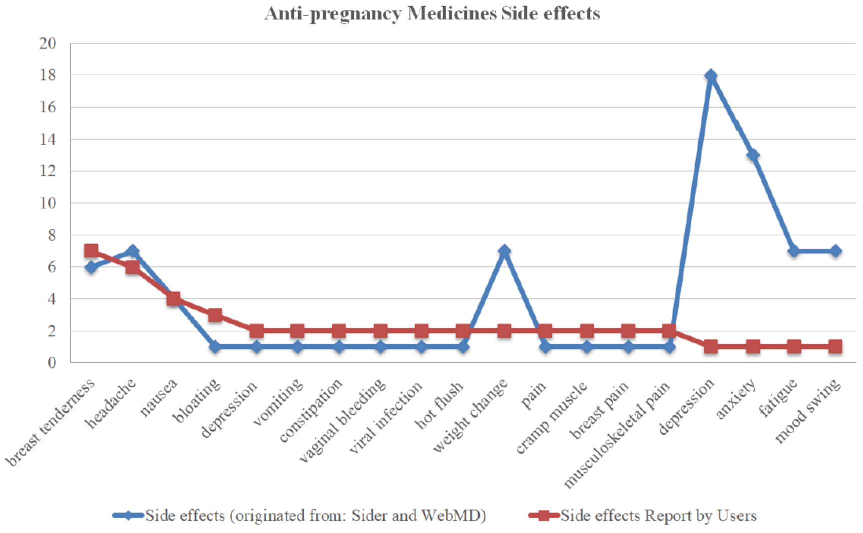


Fig. 6. Comparison of Topic Modelling of users' comments with the side effects reported on the websites of Sider and WebMD (Anti-pregnancy drugs)

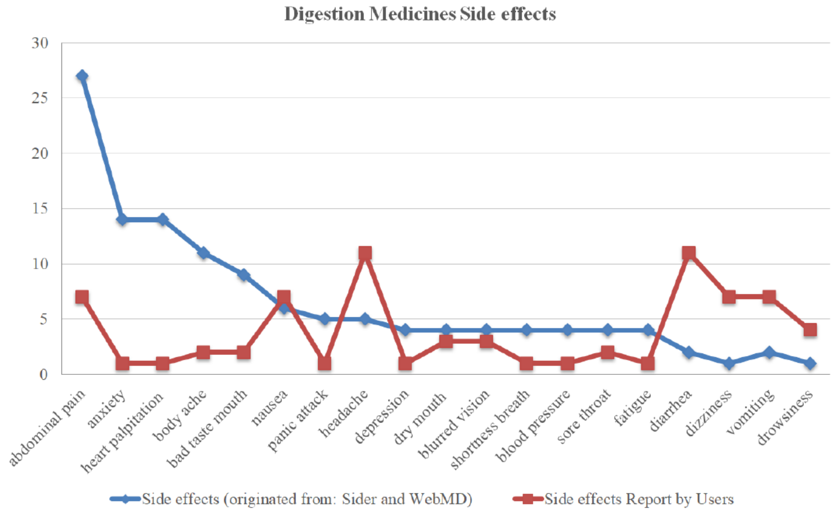


Fig. 7. Comparison of Topic Modeling of users' comments with the side effects reported in the websites of Sider and WebMD (Gastrointestinal drugs)

5. DISCUSSION

In this study, the deep learning methods of HAN and FastText were employed to classify the side effects of three classes of drugs, namely, neurotic, anti-pregnancy and gastrointestinal drugs. The reason for this investigation was high frequency of this drug consumption. Initially, the extracted data from the website "Ask a patient" were introduced to the model. And, in the pre-processing step, special characters, signs and stop words were removed, and other characters were converted into small-case letters in order to improve the text. In next phase, three classes of drugs, the side effect and the association between the former and the latter was investigated. Then, these data were exposed to classification phase (Topic Modelling) to extract 10 topics with high priority from three groups of drugs. The outputs show that the frequency of occurrence of side effects, reported in the comments in "Ask a patient" was different from that in Sider and WebMD.

Finally, the proposed model compared its output on drug's side effects with analyses of report of sites' users. The obtained results of the preliminary analysis of drug classification were presented in confusion matrices and interpreted by taking accuracy rate and false positive ratio into consideration.

In this work, it was found that Fast Text and HAN were much faster for text classification, compared to recent deep learning-based methods. We used a simple method for text classification by deep learning models. In contrast to unsupervised trained word vectors, obtained from word2vec, our word features would approximately generate appropriate sentence representations. Also, in contrast to previous studies, we suggested an end-to-end solution, based on deep learning models which do not need any handcrafted features and data pre-processing.

Our experimental findings show that each model significantly outperforms baseline methods for different datasets. Although deep neural networks, theoretically suggest higher representational power than shallow models, it is still unclear whether simple text classification would create problem or not.

6. CONCLUSION

We investigated the users' comments to identify the side effects of drugs, presented in a website, namely, "Ask a patient", then we extracted combined classification, based on three types of mostly commented diseases. Through analysis of the data with deep learning method, it was found that users' comments on side effects of drugs were biased. On the next step of this study, the comments were classified by Topic Modelling, resulting in some reports, similar to the reports published by Sider and WebMD; however, our reports had different frequency.

Our findings enable us to efficiently and quickly use large size data (batches of sample), and significantly reduce the number of updated parameters that are required for model training.

To sum up, working on publicly available data in social media opens a wide and novel window in the field of drug studies. The results of this study show that the data from social media may have noise, or may not be reliable. Accordingly, social media would be considered as a secondary source to identify side effects of drugs rather than a substitution for traditional and scientific methods of side effect identification. The proposed model in this study is capable of immediate identification of pharmacological events which most likely lead to immediate reaction and on-time discovery of these events.

Acknowledgments

The authors have no proprietary, financial, professional or other personal interest of any nature in any product, service or company. The authors alone are responsible for the content and writing of the paper. There is no conflict of interest in this article.

REFERENCES

Akhtyamova, L., Alexandrov, M., & Cardiff, J. (2017a). Adverse drug extraction in twitter data using convolutional neural network. *In, 2017 28th International Workshop on Database and Expert Systems Applications (DEXA)* (pp. 88–92). Lyon.

Akhtyamova, L., Ignatov, A., & Cardiff, J. (2017b). A Large-scale CNN ensemble for medication safety analysis. In F. Frasincar, A. Ittoo, L. Nguyen & E. Métais (Eds.) *Natural Language Processing and Information Systems*. *NLDB 2017. Lecture Notes in Computer Science* (vol. 10260, pp. 247–253). Springer, Cham.

Bordet, R., Gautier, S., Louet, H. L., Dupuis, B., & Caron, J. (2001). Analysis of the direct cost of adverse drug reactions in hospitalised patients. *European journal of clinical pharmacology, 56*(12), 935–941.

Classen, D. C., Pestotnik, S. L., Evans, R. S., Lloyd, J.F., & Burke, J. P. (1997). Adverse drug events in hospitalized patients: excess length of stay, extra costs, and attributable mortality. *Jama, 277*(4), 301–306.

Ginn, R., Pimpalkhute, P., Nikfarjam, A., Patki, A., O'Connor, K., Sarker, A., Smith, K., & Gonzalez, G. (2014). Mining Twitter for adverse drug reaction mentions, a corpus and classification benchmark. In *Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing* (pp. 1–8).

Gupta, S., Pawar, S., Ramrakhiyani, N., Palshikar, G. K., & Varma, V. (2018). Semi-supervised recurrent neural network for adverse drug reaction mention extraction. *BMC bioinformatics, 19*(8), 212.

Harpaz, R., Callahan, A., Tamang, S., Low, Y., Odgers, D., Finlayson, S., Jung, K., LePendu, P., & Shah, N. H. (2014). Text mining for adverse drug events, the promise, challenges, and state of the art. *Drug safety, 37*(10), 777–790.

Ho, T. B., Le, L., Thai, D. T., & Taewijit, S. (2016). Data-driven approach to detect and predict adverse drug reactions. *Current pharmaceutical design, 22*(23), 3498–3526.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (pp. 427–431). Association for Computational Linguistics.

Kongkaew, C., Noyce, P. R., & Ashcroft, D.M. (2008). Hospital admissions associated with adverse drug reactions: a systematic review of prospective observational studies. *Annals of Pharmacotherapy, 42*(7–8), 1017–1025.

Lee, K., Qadir, A., Hasan, S. A., Datla, V., Prakash, A., Liu, J., & Farri, O. (2017). Adverse drug event detection in tweets with semi-supervised convolutional neural networks. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 705–714). Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. doi:10.1145/3038912.3052671.

Miranda, D. S. (2018). *Automated detection of adverse drug reactions in the biomedical literature using convolutional neural networks and biomedical word embeddings.* SwissText.

Rezaei, Z., Ebrahimpour-Komleh, H., Eslami, B., Chavoshinejad, R., & Totonchi, M. (2020). Adverse Drug Reaction Detection in Social Media by Deepm Learning Methods. *Cell journal, 22*(3), 319–324.

Rison, R. A. (2013). A guide to writing case reports. *Journal of Medical Case Reports and BioMed Central Research Notes, 7*, 239. doi:10.1186/1752-1947-7-239

Sarker, A., & Gonzalez, G. (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics, 53*, 196–207.

Sultana, J., Cutroneo, P., & Trifirò, G. (2013). Clinical and economic burden of adverse drug reactions. *Journal of pharmacology, 4*(Suppl1), 73.

Tan, Y., Hu, Y., Liu, X., Yin, Z., wen Chen, X., & Liu, M. (2016). Improving drug safety, From adverse drug reaction knowledge discovery to clinical implementation. *Methods, 110*, 14–25.

Vallano, A., Cereza, G., Pedròs, C., Agustí, A., Danés, I., Aguilera, C., & Arnau, J. M. (2005). Obstacles and solutions for spontaneous reporting of adverse drug reactions in the hospital. *British journal of clinical pharmacology, 60*(6), 653–658.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics, Human Language Technologies* (pp. 1480–1489). Association for Computational Linguistics.

Applied Computer Science, vol. 16, no. 1, pp. 60–72
doi:10.23743/acs-2020-05

Submitted: 2020-03-13
Revised: 2020-03-20
Accepted: 2020-03-23

PSO, food delivery, TSP, Google maps

Sergio SOTO[*]*, Edmundo BONILLA*[*]*, Alberto PORTILLA*[**]*,*
Jose C. HERNÁNDEZ[*]*, Oscar ATRIANO*[**]*, Perfecto M. QUINTERO*[*]

FOOD DELIVERY BASED ON PSO ALGORITHM AND GOOGLE MAPS

Abstract

This article presents a solution to deal with the optimization of delivery routes problem for a mobile application focused on the restaurant sector, by using a bioinspired algorithm (PSO) to minimize delivery costs, maximize a greater number of deliveries and recommend an optional route for food delivery. Different computational experiments are carried out by using Google Maps (API) for showing the best delivery route. The results obtained are very promising for offering a good delivery service.

1. INTRODUCTION

Within the restaurant sector there are several problems and needs to be covered, but mainly there is a need for automating and managing the food distribution routes for different restaurants, which is why we are looking for a way to solve this need through technology and methodologies focused to find a solution through bioinspired algorithm. In particular, we use particle swarm optimization (PSO) that are basically focused on nature, since it is about emulating the native evolution of the species. This article presents the integration of a PSO algorithm to solve the problem of food distribution. A factor of customer loyalty is the delivery service, which contemplates the speed, that the product is not mistreated, and the temperature of the food, which are factors dependent on the efficiency with which the product is delivered prepared.

[*] Tecnológico Nacional de México, Instituto Tecnológico de Apizaco, 90300, Carretera Apizaco-Tzompantepec, Esquina Av., Instituo Tecnologico S/N, Apizaco, Tlaxcala, México, edbonn@wall.co.il
[**] Smartsoft America Business Applications S.A. de C.V. 90806, Adolfo López Mateos S/N, Texcacoac, Chiautempan, Tlaxcala, México, oatriano@smartsoftamerica.com.mx

The route planning problem is related to the problem of the traveling agent (TSP). To solve this problem, the geolocation of each person making an order to the different restaurants is taken into account. The distance from the restaurant is an important factor since a hot food order cannot take so long to be delivered.

Our approach was implemented as a module of the “Food Express” application, a mobile application focused on requesting food delivery and restaurant reservations, which was developed by Smartsoft America Business Applications S.A. of C.V. Food Express is available for Android and IOS operating systems.

The use of the application is quite simple, a user with the application on his smartphone enters the restaurant of his choice, enters the menu, selects his dish, adds a drink if desired or some type of dessert. After, the user accesses the cart and procced to the payment so that the restaurant receives the requested order. Subsequently, a delivery partner associated with the application receives a notification from the restaurant where the order to be delivered has to be collected. This way the dealer proceeds to deliver the food to the customer who requested the food from the application (Food Express, 2019).

Our proposal assumes that it is necessary to know the location of each user, since a static value is taken, that is, if the order is placed in motion the algorithm takes the longitude and latitude at the time the order is placed. These parameters are set from each user’s mobile device. Once the longitude and latitude are obtained, they agree on the Google Maps through an application programming interface (API), to calculate the trajectory of the different points to apply the optimization algorithm, then generate the best trajectory with the conditions imposed by a particle algorithm.

The problem is important since at present, most of restaurants offer home delivery service for free, the vehicle in which they make deliveries is usually an motorcycle, which has a low fuel consumption, however, the implementation of this algorithm can further save gas costs and better manage delivery personnel, covering more orders for a single person. Google maps plays an important role for this project, since it is the one that facilitates the calculation of the coordinates of each user who requests an order and in turn returns an interface where the streets are located with addresses, local, to mention some google tools maps (Google Maps, 2019).

1.1. Problem

The application of technology in the restaurant sector has had an exponential growth with the use of different technologies that exist in the market. This article intends to add intelligence, efficiency, speed and reduction of gasoline costs for the delivery of orders to different consumers. That is why through a PSO algorithm in combination with the google maps interface, the following benefits can be offered for applications in the restaurant sector:

- To improve the performance of product deliveries,
- To propose a better recruitment of personnel for the distribution,
- To reduce delivery costs,
- To have better distribution control,
- To save delivering time.

2. STATE OF THE ART

The problem associated with the delivery of food focuses on the distribution of a point A to a point B, calculating a specific delivery time. Based on distance is how you can calculate the estimated delivery time. The necessary values for this calculation are latitude and longitude to an order to provide the user with an estimated delivery time and distance. The solution is to optimize and organize the routes of different food sites, in order to minimize delivery time, reducing the number of vehicles for deliveries (Bruno, 2019; Rodriguez & Piccoli, 2020; Singh, 2020).

Google maps (API) are generally used in web pages or mobile applications and for the use of GPS (Global Positioning System). A mobile application focused on the restaurant sector that uses Google maps tools and global positioning is known as “Without Apron” (Sin Delantal Mx, 2019) A Google interface with a map concept was used to select where the order is placed for distribution (Li, Lim & Tseng, 2019). The module was developed analytics, are some of its tools used for its development.

3. TSP

The problem of the traveling agent known for the abbreviation in English TSP (Traveling Salesman Problem), basically consists of a traveler who wants to visit *n* cites, all only one each city. Starting with any of them and returning to the same place you left. This problem can solve real-life situations that can be formulated differently. The problem of the traveling agent within the branches that helps solve problems within robotics, mechanics, automotive industry, logistics mainly focused on optimization problems (Rodríguez & Ruiz, 2012; Stockdale 2011; Archetti, Feillet, Mor & Speranza, 2020).

4. BACKGROUND

4.1. PSO

Within the bioinspired algorithm there is a very busy one for the route optimization known as PSO (Particle Swarm Optimization). This methodology arises when developed by Kennedy and Eberhart (1995), one of the hypotheses of each particle or agent that represents bees, ants or birds or some individual who is in a social group with a guided search, the particle the best solution you have found so far fulfills the correct role as leader.

The idea of the proposed algorithm is to obtain a particle solution that evolves in order to find a better solution in its path. Within the literature, the theoretical foundations of this method are the movement of each of the existing particles focused on a common objective which is conditioned by two factors: the fists are the nostalgia or autobiographical memory of the particle and the social influence from the group or the swarm (Fontana, 2004).

Each particle has an instantaneous position of the population in the N-dimensional space represented within the domain of the objective function that is proposed for a possible solution, *N* is the number of unknowns of the problem. In addition, the evolutionary process is diminished by moving each particle within the solution space with a speed that will be transformed according to its current speed, the particle memory and global information within the entire swarm.

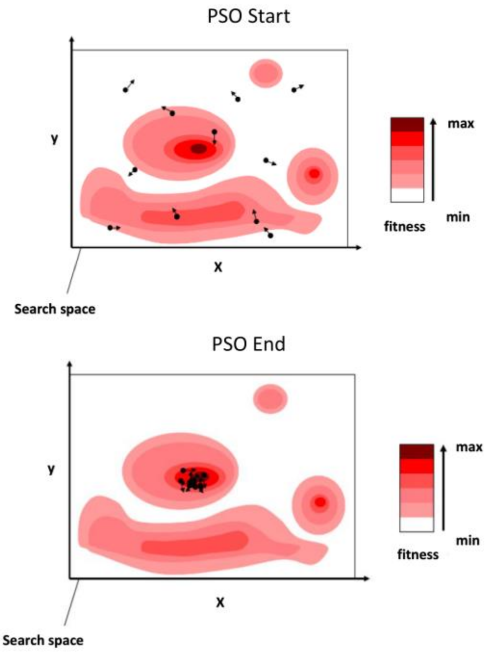


Fig. 1. Example of algorithm PSO start and end (Di Caro, 2012)

Within the search space, the particle is optimized using the fitness solution, which is the best solution space within the search space (Figure 1). At the beginning of the algorithm, the particles remain distributed in space, once the algorithm begins to run, the particles find and individual and group fitness solution to find a point where the solution is the most optimal for the particles.

The PSO consists of an iterative and stochastic process that operates on a cluster of particles. The position of each particles represents a potential solution to the problem that is begin solved. Regularly, a particle Φ_i is composed of three vectors each with two velocity values or well-known as fitness. The following values make up fitness:

- The vector $X_i = \{X_i1, X_i2,\ldots, X_in\}$ stores the current position of the particle in space. The size of this vector depends on the number of variables needed to solve the problem.
- The vector Φ $Best_i = \{\Phi_i1, \Phi_i2,\ldots ,\Phi_in\}$ in stores the position of the best solution found by the particle so far.
- The velocity vector $V_i = \{V_i1, V_i2, \ldots, V_in\}$ stores the direction according to which the particle will move.
- The *fitness* X_i fitness value stores the agreement value of the current solution (vector X_i).
- The fitness *fitness* value Φ $Best_i$ stores the agreement value of the best local solution found so far (vector $Best_i$).

5. EXPERIMENTS

A software was implemented for the optimization of food delivery routes for the restaurant sector, applying the PSO bioinspired algorithm which will benefit food deliveries optimizing time, gasoline costs and better delivery management, in addition, giving solution to the problem of the traveling agent which merges with the incorporation of Google maps tools, to have a pleasant visual interface for the end user.

The algorithm has the task of being better than the calculation of matrices of Google maps, working more efficiently and intelligently (Zhou et al., 2019).

The first activity is to carry out a conceptual design to be able to interpret the needs of the restaurant sector, to understand in an illustrated way the operation of the algorithm, as well as to solve the problem of the traveling agent, giving efficiency to the delivery. Then the geolocation of the client is needed to obtain its latitude and longitude, where food delivery will be made, that information reaches the application in which the orders for food delivery are generated.

Once the location is obtained, the web service is activated, where you will obtain the request of the different delivery points, obtaining the geolocation parameters of the users activates the PSO algorithm, where you find the best delivery route for the restaurant. The restaurant obtains a recommendation of the algorithm within the Google maps interface, so that it visualizes its possible distribution trajectory. Figure 2 shows the possible solution through the PSO algorithm, using Google maps tools and operation of requests and interactions with all users.

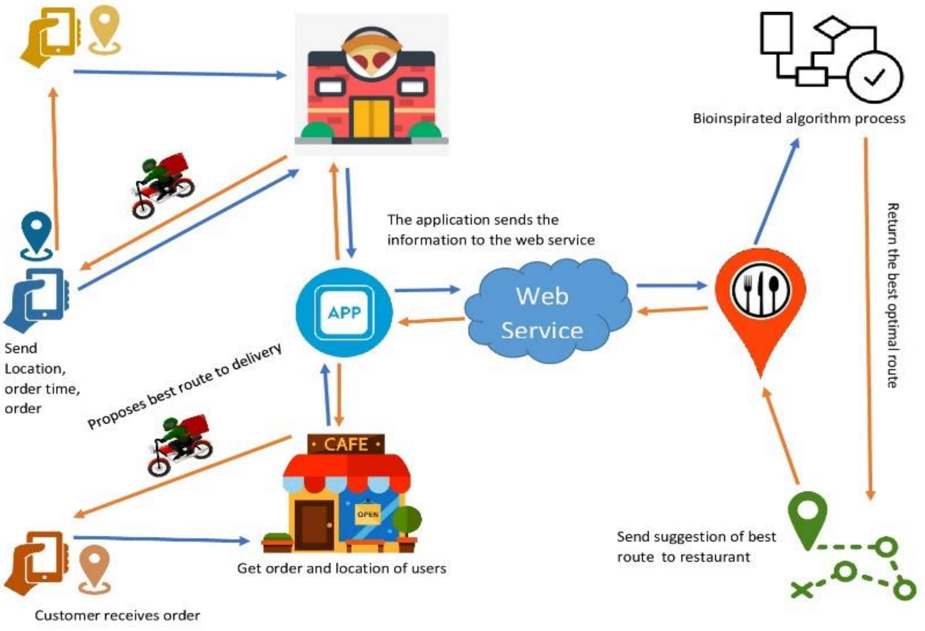
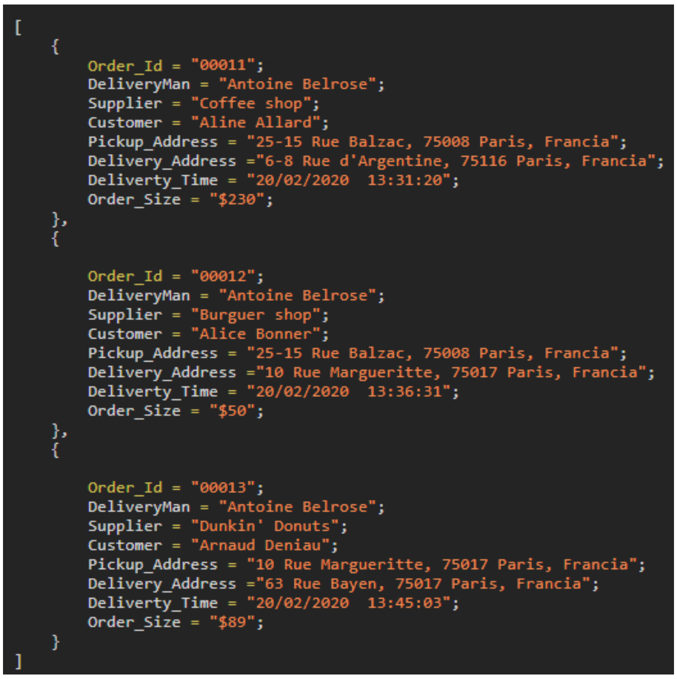

Bioinspirated algorithm process
The application sends the information to the web service
Return the best optimal route
APP
Web Service
Send Location, order time, order
Proposes best route to delivery
CAFE
OPEN
Send suggestion of best route to restaurant
Get order and location of users
Customer receives order

Fig. 2. Conceptual design of problem solving based on food deliveries

Once the conceptual design was established, the next step was to develop the interface where the global map of Google maps (API) is displayed, showing the location of the different delivery points. Executing the PSO algorithm with the delivery points yields a delivery recommendation, calculating the delivery time, algorithm iterations and delivery points. Some parameters for the algorithm have been established such as the size of the particles, the position of the particles, speed, the fitness particles to be able to execute the algorithm correctly (Chen, 2019; Paciarotti, Bevilacqua, Ciarapica, Mazzuto & Postacchini, 2019). Once the parameters are taken, the map is plotted on the map with the route proposed by the algorithm and the API.

The web service provides the opportunity to have better operability of data in a short amount of time, which provides the best execution of the algorithm. The PSO has the advantage of working with few elements, in this way the execution of the web service gives us the necessary speed to obtain data and the operations process. Using a protocol type SOPA (Simple Object Access Protocol). To give an example of some data that the web service works in combination with the algorithm, the result of an order 3 deliveries by the same distributor is presented, with different orders located in Paris near the triumphal arch.

Where it contains data that serve to understand the order. The set of data that you save by distribution order, some of the data are: order number, delivery person, place of collection, customer, supplier, place of delivery, delivery time and cost of the order. With which the latitude and longitude can be obtained from the address with Google API methods (Figure 3). The web service returns an object or type Json with the values of the attributes already mentioned, as part of an array where the problem is already optimized.

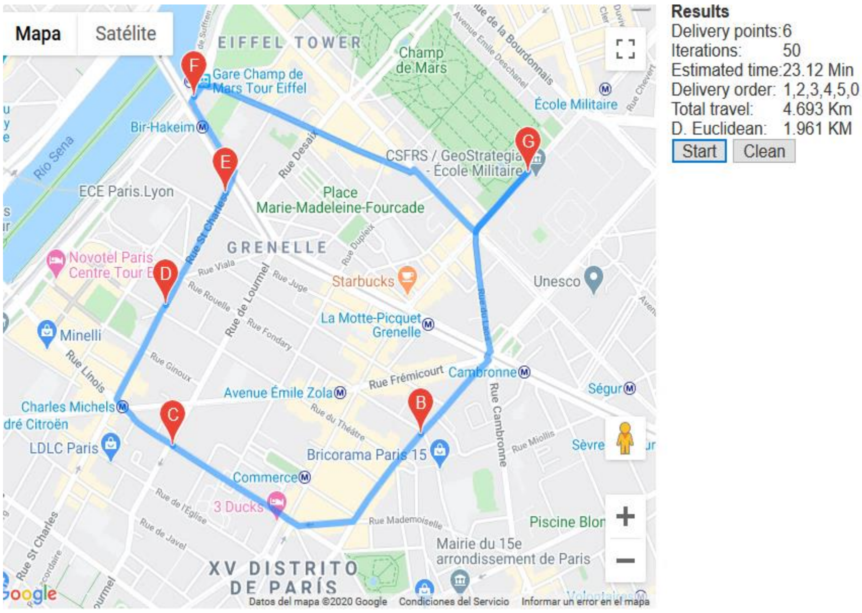


```
[
    {
        Order_Id = "00011";
        DeliveryMan = "Antoine Belrose";
        Supplier = "Coffee shop";
        Customer = "Aline Allard";
        Pickup_Address = "25-15 Rue Balzac, 75008 Paris, Francia";
        Delivery_Address ="6-8 Rue d'Argentine, 75116 Paris, Francia";
        Deliverty_Time = "20/02/2020  13:31:20";
        Order_Size = "$230";
    },
    {

        Order_Id = "00012";
        DeliveryMan = "Antoine Belrose";
        Supplier = "Burguer shop";
        Customer = "Alice Bonner";
        Pickup_Address = "25-15 Rue Balzac, 75008 Paris, Francia";
        Delivery_Address ="10 Rue Margueritte, 75017 Paris, Francia";
        Deliverty_Time = "20/02/2020  13:36:31";
        Order_Size = "$50";
    },
    {

        Order_Id = "00013";
        DeliveryMan = "Antoine Belrose";
        Supplier = "Dunkin' Donuts";
        Customer = "Arnaud Deniau";
        Pickup_Address = "10 Rue Margueritte, 75017 Paris, Francia";
        Delivery_Address ="63 Rue Bayen, 75017 Paris, Francia";
        Deliverty_Time = "20/02/2020  13:45:03";
        Order_Size = "$89";
    }
]
```

Fig. 3. Web service results.

Showing some small results that are delivery points, algorithm iterations, estimated travel time, delivery order based on orders to be delivery, total travel in kilometers and a proposed route for food delivery. Initially with him experiment a delivery with 5 delivery points was analyzed, using an iteration of 50 to find the best optimized path. The algorithm handles a particle speed that in this case is 0.1, with a solution space of 50. By applying the algorithm on the map, it generates a delivery recommendation to the dealer (Figure 4).

A test was carried out in the city of Paris, near the Eiffel tower within the fifteenth district in order to show the functionality of the algorithm in another city that is not Mexico.



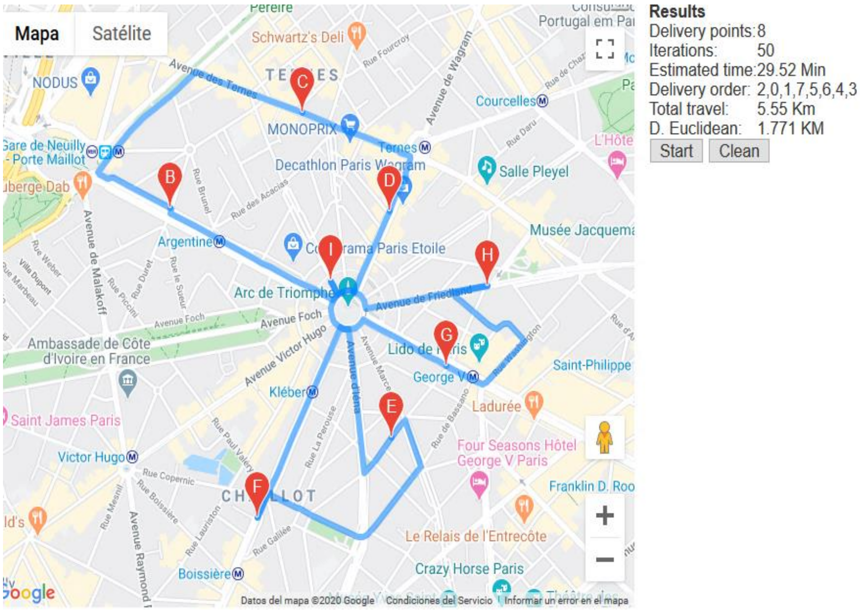
Results
Delivery points:6
Iterations: 50
Estimated time:23.12 Min
Delivery order: 1,2,3,4,5,0
Total travel: 4.693 Km
D. Euclidean: 1.961 KM

Fig. 4. Execution of the route optimizer with PSO algorithm (5 delivery points in Paris)

Six delivery points are shown because one is from the starting point, where the restaurant is located, that is why there has to be a return to the starting point. The path is drawn by Google library, the marks within the map are used by same library.

The following test was performed with 7 delivery points, 50 iterations in the search space to find the best optimal route. Based on the algorithm, you need a particle velocity that used a velocity of 0.1 for this test. One of the requirements to cover for this problem was to know the total distance of the route, this measurement is calculated using the API matrix, calculates the route on streets or the route of the distribution between the streets (Figure 5).

This test was carried out in the same way in Paris near the triumphal arch. A good delivery time was observed for the distribution points, since it did not exceed 30 minutes of distribution.



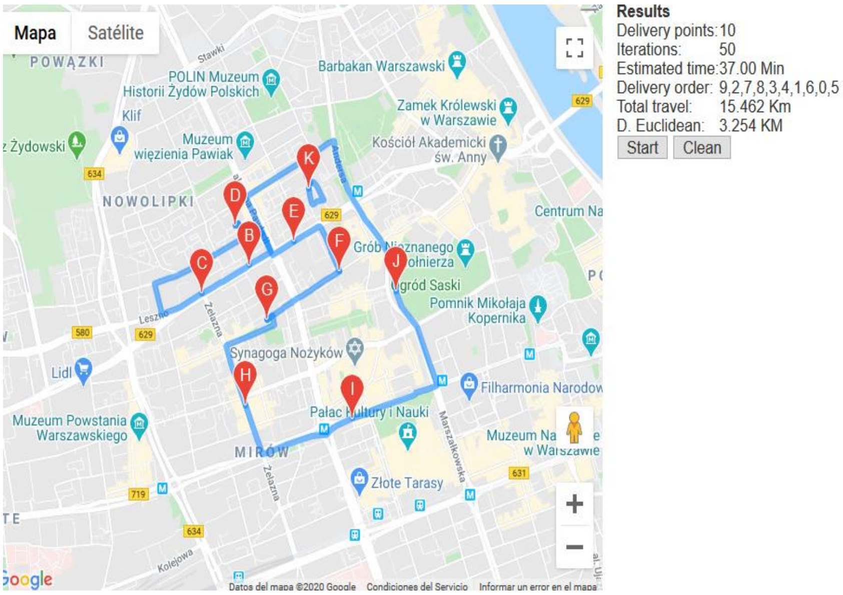
```
Results
Delivery points:8
Iterations:      50
Estimated time:29.52 Min
Delivery order: 2,0,1,7,5,6,4,3
Total travel:    5.55 Km
D. Euclidean:    1.771 KM
Start  Clean
```

Fig. 5. Execution of the route optimizer with PSO algorithm (7 delivery points in Paris)

Next, the test is presented with a delivery of 9 orders, with 50 iterations at a speed of 0.1, since it is recommended in the literature. This test shows a greater number of deliveries, which is reflected in the estimated delivery time (Figure 6), so this number of orders is recommended as long as the orders are near the restaurant. Contemplating that they are meals that are served at a hot temperature, otherwise the different orders can be delivered in s slightly larger range than to the restaurant area.

A test presented with 9 delivery points shows that the delivery time can take more than 35 minutes, which up to this point the algorithm begins to have limitations for hot food deliveries. Even delivery points are feasible for cold dishes. This test was conducted in the capital of Poland, Warsaw. In order to demonstrate the functionality of the PSO algorithm, to visualize the effectiveness of the application.

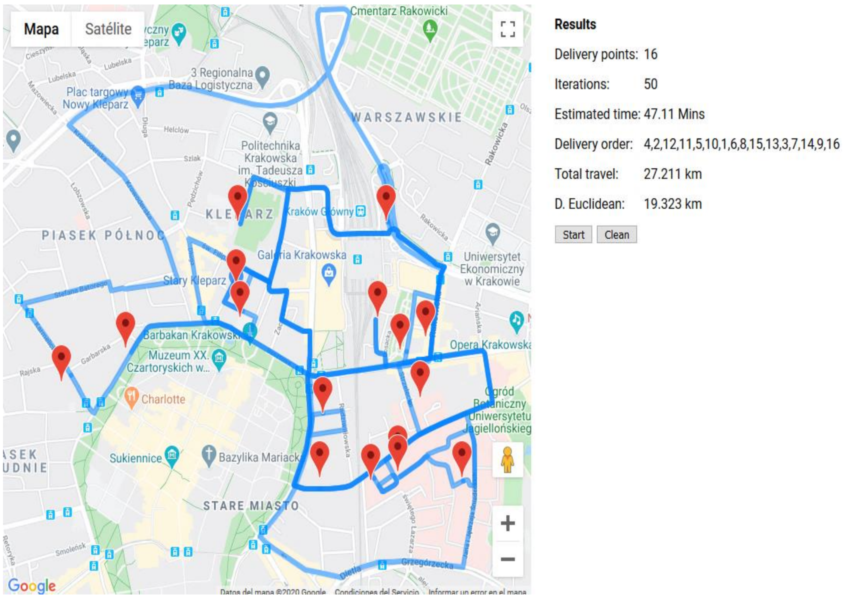
To test the effectiveness of the PSO algorithm, an example was used with 15 delivery points within a radius of no more than 10 kilometers where the restaurant is located. In this case 10 kilometers of radius are not exceeded, remember that 16 points are displayed on the map (Figure 7) since one of them is the restaurant, where you have to leave and return. Once having the results, the delivery time that exceeds 47 minutes of delivery was analyzed, with a journey of 27 km, in this way you have an approximate of how much fuel the distributor will spend, in this way they are better managed expenses and reducing costs.



Results
Delivery points: 10
Iterations: 50
Estimated time: 37.00 Min
Delivery order: 9,2,7,8,3,4,1,6,0,5
Total travel: 15.462 Km
D. Euclidean: 3.254 KM
Start Clean

Fig. 6. Execution of the route optimizer with PSO algorithm (9 delivery points in Warsaw)

For this test location of Cracow was used, in order to understand the behavior of the algorithm with 15 delivery points. The more delivery points the dealer has, the road begins to be longer and with more total travel time, at this point the algorithm begins to increase exponentially over time.


Mapa
Satélite
Cmentarz Rakowicki
3 Regionalna Baza Logistyczna
Plac targowy Nowy Kleparz
WARSZAWSKIE
Politechnika Krakowska im. Tadeusza Kościuszki
KLEPARZ
Kraków Główny
PIASEK PÓŁNOC
Galeria Krakowska
Uniwersytet Ekonomiczny w Krakowie
Stary Kleparz
Barbakan Krakowski
Opera Krakowska
Muzeum XX. Czartoryskich w...
Charlotte
Ogród Botaniczny Uniwersytetu Jagiellońskiego
Sukiennice
Bazylika Mariacka
PIASEK... UDNIE
STARE MIASTO
Google
Datos del mapa ©2020 Google
Condiciones del Servicio
Informar un error en el mapa

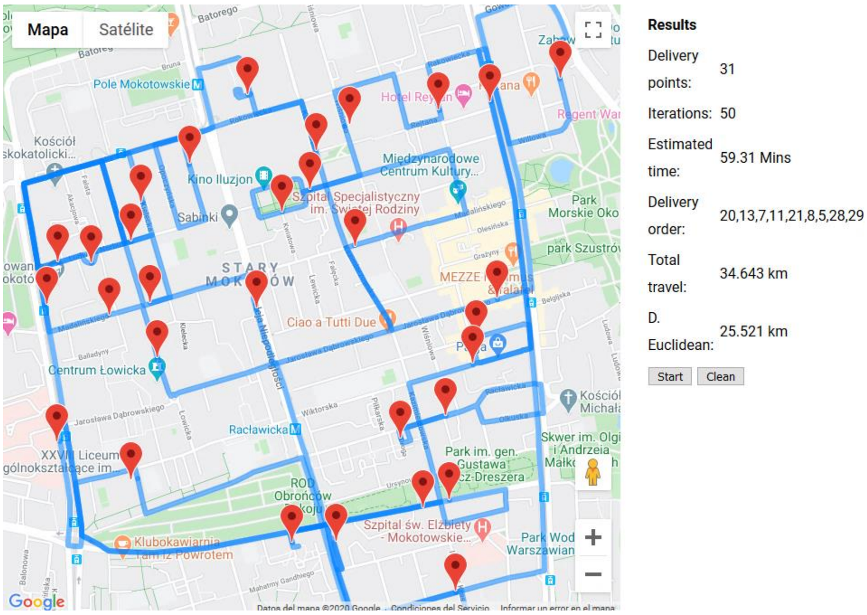
Results
Delivery points: 16
Iterations: 50
Estimated time: 47.11 Mins
Delivery order: 4,2,12,11,5,10,1,6,8,15,13,3,7,14,9,16
Total travel: 27.211 km
D. Euclidean: 19.323 km
Start Clean

Fig. 7. Execution of the route optimizer with PSO algorithm (15 delivery points in Cracow)

Finally, an experiment was carried out with a number of 30 deliveries, at this point the deliveries already meet a long time, for the number of deliveries by a single deliveryman, in this case being an optimization for food service is too much delivery time. This number of deliveries for food is not recommended, it can be functional for another sector such as courier, logistics, to name a few. In this test the parameters of 50 iterations were taken with a speed similar to the previous tests of 0.1, with a search space of 30 particles.

The test was carried out in order to demonstrate a delivery time, to know an approximate route for the deliveryman. A possible delivery is presented on the map (Figure 8), but this may vary depending on unexpected streets of blockages.

Mokotow was the place where this test took place, it is the city with the greatest number of inhabitants of the city of Warsaw. Performing a test where 30 deliveries are made in less than an hour. Walking a total of 34 km, which would have to be done by a single distributed. Until this the algorithm fulfills its function of optimizing, however, the estimated delivery time is no longer feasible. Therefore, this number of deliveries is no longer efficient.


Mapa
Satélite
Batorego
Pole Mokotowskie
Kościół
skokatolicki...
Kino Iluzjon
Sabinki
Hotel Rey...n
Międzynarodowe Centrum Kultury...
Szpital Specjalistyczny im. Świętej Rodziny
STARY MOKOTÓW
Park Morskie Oko
park Szustra
MEZZE
Ciao a Tutti Due
Centrum Łowicka
Racławicka
XXVII Liceum
ROD Obrońców
Klubokawiarnia Tam i z Powrotem
Szpital św. Elżbiety - Mokotowskie...
Park im. gen. Gustawa ...cz-Dreszera
Skwer im. Olgi i Andrzeja Małk...
Kościół Michała
Park Wodny Warszawian...
Google
Datos del mapa ©2020 Google Condiciones del Servicio Informar un error en el mapa
Start
Clean

Results

Delivery points: 31

Iterations: 50

Estimated time: 59.31 Mins

Delivery order: 20,13,7,11,21,8,5,28,29

Total travel: 34.643 km

D. Euclidean: 25.521 km

Fig. 8. Execution of the route optimizer with PSO algorithm (30 delivery points in Mokotow)

In addition, it should be mentioned that the execution time can vary if certain values that make the process more robust are modified, for example, if the particle size is change to a larger amount it may take a little longer since this algorithm is usually exponential.

6. CONCLUSION

This article proposes the development of a module for a mobile application in order to optimize food delivery routes to the restaurant sector through a particle swarm optimization algorithm, using the Google maps API. In order to propose restaurants a route where they can make different deliveries of orders in a single trip by delivery. Within the results obtained, the good functionality of the module can be concluded, the efficiency, speed of the algorithm can be affected in the device in which the PSO is executed.

Similarly, the benefit of this implementation seeks the adaptability of the different restaurants that offer home delivery service. The PSO algorithm can be exploited and overcome with implementations that seek an improvement of the algorithm through combinations of other methods such as fuzzy logic type I. Future work is the implementation of fuzzy logic and the PSO algorithm applying a natural language, giving each restaurant variables of how feasible the delivery to certain addresses and which is not as feasible depending on the distance and delivery points.

REFERENCES

Archetti, C., Feillet, D., Mor, A., & Speranza, M. G. (2020). Dynamic traveling salesman problem with stochastic release dates. *European Journal of Operational Research*, *280*(3), 832–844.

Bruno, L. (2019). *Solving a food-delivery problem with a Vehicle Routing Problem-based approach* (Doctoral dissertation). Politecnico di Torino, Torino.

Chen, L. W. (2019). Impact Assessment of Food Delivery on Urban Traffic. In *2019 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)* (pp. 236–241). Zhengzhou, China: IEEE.

Di Caro, G.A. (2012) Collective and Swarm Intelligence. Retrieved from https://sci2s.ugr.es/sites/default/files/files/Teaching/GraduatesCourses/Metaheuristicas

Fontana, R. J. (2004). Recent system applications of short-pulse ultra-wideband (uwb) technology. *IEEE Transactions on microwave theory and techniques*, *52*(9), 2087-2104.

Food Express. (2019). Application for food delivery. Retrieved from https://www.foodexpress.com.mx/index.xhtml

Google Maps. (2019). Satellite map display application. Retrieved from https://developers.google.com/maps/documentation

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. In *Proceeding of ICNN'95 – International Conference on Neural Networks* (vol. 4, pp. 1942–1948). IEEE.

Li, Y., Lim, M. K., & Tseng, M. L. (2019). A green vehicle routing model based on modified particle swarm optimization for cold chain logistics. *Industrial Management & Data Systems*, 119(3), 473–494. doi:10.1108/IMDS-07-2018-0314

Paciarotti, C., Bevilacqua, M., Ciarapica, F. E., Mazzuto, G., & Postacchini, L. (2019). An efficiency analysis of food distribution system through data envelopment analysis. *International Journal of Operational Research*, *36*(4), 538–554.

Rodríguez, A., & Ruiz, R. (2012). The effect of the asymmetry of road transportation networks on the traveling salesman problem. *Computers & Operations Research*, *39*(7), 1566–1576.

Rodriguez, J., & Piccoli, G. (2020). Seeking Competitive Advantage Through Platform-Enabled Resources: The Case of Food Delivery Platforms. *In Proceedings of the 53rd Hawaii International Conference on System Sciences* (pp. 5545-5554).

Sin Delantal Mx. (2019). Application for food delivery. Retrieved from https://www.sindelantal.mx

Singh, G. (2020). Online Food Delivery Services: A Study on Demographic Attributes. *Our Heritage*, *68*(1), 2147-2165.

Stockdale, M. L. (2011). *El problema del agente viajero: un algoritmo heurístico y una aplicación*. Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires.

Zhou, H., Yao, P., Xiao, Y., Fan, K., Zhang, Z., Gong, T., Zhao, L., Deng, M., Liu, C., & Ling, P. (2019). Friction and wear maps of copper metal matrix composites with different iron volume content. *Tribology International*, *132*, 199-210.

Applied Computer Science, vol. 16, no. 1, pp. 73–84
doi:10.23743/acs-2020-06

Submitted: 2020-03-05
Revised: 2020-03-07
Accepted: 2020-03-13

recognition software, car number plate, design

Mohanad ABDULHAMID[*], *Njagi KINYUA*[**]

SOFTWARE FOR RECOGNITION OF CAR NUMBER PLATE

Abstract

The purpose of this paper is to design and implement an automatic number plate recognition system. The system has still images as the input, and extracts a string corresponding to the plate number, which is used to obtain the output user data from a suitable database. The system extracts data from a license plate and automatically reads it with no prior assumption of background made. License plate extraction is based on plate features, such as texture, and all characters segmented from the plate are passed individually to a character recognition stage for reading. The string output is then used to query a relational database to obtain the desired user data. This particular paper utilizes the intersection of a hat filtered image and a texture mask as the means of locating the number plate within the image. The accuracy of location of the number plate with an image set of 100 images is 68%.

1. INTRODUCTION

Automatic Number Plate Recognition (ANPR) is a mass surveillance method that uses optical character recognition on images to read the license plate on vehicles using existing closed circuit television cameras or road-rule enforcement cameras, or ones specifically designed for the task-some systems commonly use infrared lighting to allow the camera to take the picture at any time of day. They are used for various tasks, including electronic toll collection on pay per use roads, restricted car identification access control schemes such as for pay parking-lots or for secured office compounds, monitoring traffic activity such as red light adherence in an intersection and for direct marketing. ANPR technology tends to be region specific, owing to plate variation from place to place. The first ANPR was invented in 1976 at the Police Scientific Development Branch in the UK.

[*] Al-Hikma University, Karada Kharidge, Baghdad, Iraq, moh1hamid@yahoo.com
[**] University of Nairobi, P.O. Box 30197-00100, Nairobi, Kenya, researcher12018@yahoo.com

Prototype systems were working by 1979 and contracts were let to produce industrial systems at the Computer Recognition Systems (CRS) in Wokingham, UK. Early trial systems were deployed on the A1 road and at the Dartford Tunnel (Badr & Abdelwahab, 2011; Sharma, 2018).

License Plate Recognition(LPR) has a wide range of applications, which use the extracted plate number and optional images to create automated solutions for various problems (Lin, Lin & Liu, 2018; Xie & Zhang, 2018), these include:

1. Parking – the plate number is used to automatically enter pre-paid members and calculate parking fee for non-members by comparing the exit and entry times.
2. Access control – a gate automatically opens for authorized members in a secured area, thus replacing or assisting the security guard. The events are logged on a database and could be used to search the history of events
3. Law enforcement – the plate number is used to produce a violation fine on speed or red-light systems. The manual process of preparing a violation fine is replaced by an automated process which reduces the overhead and turnaround time. Using a list of stolen cars or unpaid fines, an Automatic License Plate Recognition(ALPR) system deployed on the roadside, may perform a real-time match between the passing cars and those on the list, hence raise alerts for passing cars on the list.
4. Marketing tool – the car plates may be used to compile a list of frequent visitors for marketing purposes, or to build traffic profiles, such as the frequency of entry verses the hour or day.
5. Tolling – the car number is used to calculate the travel fee in a toll-road, or used to double-check the ticket.

Such routine tasks, possibly handling large volumes of traffic, maybe more efficiently done if automated than if done by humans. This would reduce associated costs of operation, increase processing speed and minimize errors that may result due to fatigue or monotony if a human operator were involved (Silva & Jung, 2018).

The automatic plate recognition system may be decomposed into three blocks: an image processing stage, an optical character recognition stage and the database.

2. PROBLEM SOLUTION AND DESIGN

The automatic license plate recognition application is developed using the license plate extraction, character segmentation and optical character recognition modules. The automatic number license plate presented in this work is developed in MATLAB 7.6. Implementation is grouped into image processing, optical character recognition and database blocks.

2.1. Image processing block

This block receives input images of the vehicles and produces cropped images which are passed to the optical character recognition block that succeeds it. The image processing block is based on image pre-processing and segmentation steps. First, the input color (RGB) images are converted into grayscale images. An RGB image, sometimes referred to as a *true-color* image, is stored as an m-by-n-by-3 data array that defines red, green, and blue color components for each individual pixel. The color of each pixel is determined by the combination of the red, green, and blue intensities stored in each color plane at the pixel's location. A grayscale image is a data matrix whose values represent intensities within some range.

This conversion allows ease of thresholding, which creates a binary image using a certain threshold. Adaptive thresholding using the Otsu thresholding scheme; which assigns pixels with grey-level above a threshold (chosen so as to minimize the intra-class variance of the black and white pixels) in the grayscale scale image are assigned binary value “1” in the binary image, while those below are assigned binary value “0” are implemented using the inbuilt *graythresh* function.

Next, morphological operations opening and closing are used. The morphological closing operation is carried out using a rectangular structuring element of a size much greater than the inter-character spacing of the license plate characters, resulting in blurring of the plate characters in the license plate. Subtracting the resultant image from the original thresholded image, which constitutes top-hat filtering, removes large parts of constant intensity background, leaving plate characters and other fine details in the image intact. A morphological opening operation with a structuring element, of width less than the inter- character spacing, is then used to remove the unimportant fine details without affecting the characters themselves.

An image mask which isolates the regions with the largest difference in pixel intensities in a given neighborhood is then developed using a range filter. The function *rangefilt* is used to obtain an output array of same size as the input image, where each output pixel contains the range value, as the difference between maximum and minimum pixel intensity values in the 3-by-3 neighborhood around the corresponding pixel of the input image. The range filtered image is then binarized using an experimentally determined threshold of 0.4, which preserved as much license plate detail as possible, without including too many noise objects. A morphological closing operation is then done on the binary image with a 9-pixel, square structuring element to remove thin objects in image due to local transitions at the edges of the input image. Larger structuring elements merge objects, yielding large area masks that left many noise objects when overlaid with the top-hat filtered image, whereas smaller structuring elements yield fragmented characters. An area opening is then used to remove from the binary image all connected objects that had fewer than 300 pixels, the figure being determined by trial and error, to produce the binary masking image.

The image created by over-laying the top-hat filtered image with the mask by carrying out the logical AND of the two images is then labeled using the *bwlabel* function. The *bwlabel* function is used to search for connected components and labels them with unique numbers; it takes a binary input image and a value specifying the connectivity of objects. The function returns a matrix *L* of the same size as the input image, containing labels for the connected objects in that input image, and the number of connected objects found in the input image. The elements of *L* are integer values greater than or equal to 0.

The *regionprops* function is then used to measure object or region properties in an image and returns them in a structure array. When applied to an image with labeled components, it creates one structure element for each component; with the structure array having *area, centroid and bounding box* fields. The *area* field is scalar representing the actual number of pixels in the region. The *bounding box* field is a vector representing the smallest rectangle containing the region defined by co-ordinates of the upper left corner of the bounding box and the width of the bounding box along each dimension, in length and height. The *centroid* field is a vector that specifies the center of mass of the region; the first element of centroid is the horizontal coordinate (or x-coordinate) of the center of mass, and the second element is the vertical coordinate (or y-coordinate).

The *centroid* field of the *regionprops* function is used to determine the vertical coordinate of all labeled objects obtained from the logical AND of the top-hat filtered image and the texture map. Since few objects are members of either set, resultant intersection of both sets is composed largely of license plate characters. The y-coordinate values corresponding to the objects are observed to be randomly distributed over the image height, but objects corresponding to license plate characters had approximately the same centroid value. This variation in the centroid value is exhibited in the ones digit of the value; hence a rounding-off of the centroid values to the nearest ten-arrived at by trial and error eliminates this variation. The mode of the rounded vertical coordinates of objects in this intersection is thus the average height of the license plate characters on the image. Selecting all objects intersecting with a horizontal line, running across the image at this modal height yields the license plate characters. In very few cases, noise objects are included. The *area* field of the *regionprops* function is used to remove all noise objects with an area of less than 100 pixels, determined as the minimum area of license plate character-corresponding to the numeral 1.

Further filtering of noise objects requires the orientation of the major axis of the ellipse that has the same variance as the region of each object lie between 45 to 90 degrees from the horizontal axis. This allows removal of horizontally aligned noise objects resulting from dilation of fragmented license plate edges due to the closing operation. The Euler number of each object is also used to remove noise objects. The Euler number is a measure of the topology of an image. It is defined as the total number of objects in the image minus the number of holes in those objects. Character (B) has Euler number of -1, indicating that the number of holes is greater than the number of objects; characters such as (A, O, P, D, Q) have an Euler number of zero, whereas all other simple characters have an Euler number of 1. Due to nails in the license plate, noise is introduced to character objects due to the connectivity used in labeling. An object having an Euler number of 1 in isolation would show an Euler number of two if such a nail object is introduced in its bounding box due to connectivity. As a result, all objects with an Euler number greater than 2 or less than -1 are considered to be noise objects and are removed. The remaining objects are subsequently considered to be characters and are passed to the optical character recognition block.

The *bounding box* field of the *regionprops* function is then used to obtain the smallest rectangles containing each object in the resultant image. These rectangles are then passed to the *imcrop* function which crops the required labeled objects corresponding to the license plate characters in the image, based on the rectangles' top-left coordinates, their widths, and heights respectively. It is these objects that are passed to the optical character recognition block. The quality of segmentation is strongly related to the choice of the structuring element's size on the plate enhancement phase, choice of the threshold used for image binarization, the relative angle between camera and plate and the quality of the image.

2.2. Optical character recognition block

2.2.1. Neural network classifier approach

A back-propagation neural network classifier is first adopted and is trained on thirty four vectors; corresponding to all possible characters; each having 150 elements; corresponding to image objects of resolution 15×10. Each target vector is a 34-element vector with a 1 in the position of the letter it represents, and 0's everywhere else. For example, the letter A is to be represented by a 1 in the first element (as A is the first letter of the alphabet), and 0's in elements two through thirty four. The network receives the 150 Boolean values as a 150-element input vector. It is then required to identify the letter by responding with a 34-element output vector. The 34 elements of the output vector each represent a letter. To operate correctly, the network would respond with a 1 in the position of the letter being presented to the network. All other values in the output vector would be 0. The neural network thus has 150 inputs and 34 neurons in its output layer to identify the letters. The network is a two-layer log-sigmoid network, with 10 neurons in the hidden layer. Training is done for 5000 iterations, for gradient descent with momentum. Gradient descent with momentum, implemented by the *traingdm* function, allows a network to respond not only to the local gradient, but also to recent trends in the error surface. Momentum allows the network to ignore small features in the error surface, hence slide through shallow local minima. Without momentum a network can get stuck in such a minimum.

While the neural network classifier is usually preferred owing to its higher cognitive rate and its ability to give reasonable results when presented with new object which it has not been trained on, several shortcomings result in adoption of the template matching approach. With character recognition for license plate applications, the neural network classifier requires frequent retraining, preferably for each car image due to the large input vector size, having 150 binary elements for each character.

Accuracy improves as the number of neurons in the hidden layer and as size of input training vector increases. Increasing number of neurons in the hidden layer results in a speed penalty while increasing size of the input training vector is limited by memory constraints. Training requires either 5000 iterations or a steady-state error of less than 0.1, which make it slower than the template matching approach. In this case, the neural network classifier has a poor cognitive rate and misclassified well segmented and obvious characters.

2.2.2. The template matching approach

The template matching approach is then implemented. The templates used have a resolution of 42×24, hence rescaling the license plate objects prior to template matching is necessary. Character recognition is based on calculating the correlation metric, implemented using the *corr2* function, which computes the correlation coefficient between two matrices of the same size. All images from extracted objects and the template set are thus represented as 42×24 intensity matrices. The template images corresponding to the 34 possible characters A to Z and 0 to 9 are saved, and template matching is implemented by using the correlation between each extracted object from the image against all the images in the template. This removes the presumption that all license plates begin with letters (which would affect recognition of diplomatic license plates) or that license plate begins with letter K (which would affect recognition of foreign license plates). The correlation coefficients for each extracted object with the template set is ordered into a 34 element array, and the index of the element having the highest correlation coefficient used to identify the corresponding similarly indexed character. The characters corresponding to each extracted object are then concatenated to form a string, which is the detected vehicle registration number.

2.3. Database implementation

For a car identification access system, a database is needed. Since the basic segmentation and character recognition modules have been implemented, the output string from these modules is to be used to query a test database and extract hypothetical user data. While the original idea is to use a MYSQL database and use a PHP script to query the database, this is found to be unnecessary. MATLAB supports database queries from most databases since it implements Java Database

Connector (JDBC), and supports JDBC to Object Database Connector(ODBC) inter-conversion. This database support is implemented by the MATLAB database toolbox which requires that a data source should be set up first. The data source consists of data that the toolbox accesses and information required to find the data, such as driver, directory, server, or network names. Data sources interact with ODBC drivers or JDBC drivers. An ODBC driver is a standard windows interface that enables communication between database management systems and SQL-based applications. A JDBC driver is a standard interface that enables communication between applications based on Java and database management systems. The database toolbox software is based on Java. It uses a JDBC/ODBC bridge to connect to the ODBC driver of a database, which is automatically installed as part of the MATLAB Java Virtual Machine (JVM).

Thus a Microsoft access data source is set up, and a test database is created, having three fields; the car registration number, the car's color and its making or model; with the registration number field as the primary key. A MATLAB script is then used to connect to this database using the JDBC/ODBC driver bridge to obtain a connection object which is then used to pass SQL queries to the test database. In order to access the database, first a connection to the database is created using the database function. This function takes data source, username and password as arguments. The data source is the name of the data source in the ODBC for Windows. This has to be configured before running MATLAB. The username of the user has on the database to be accessed. The password is given with the defined username to access the database. These arguments are passed to the database function as strings.

Once the connection is established, queries to the database are performed using the fetch function. This function takes a connection pointer and an SQL query string as arguments. The connection pointer is the connection created with the database function. The SQL query string contains the desired SQL query. The queries used are based on the license number string, passed from the character recognition block. User data corresponding to the predefined fields is thus retrieved.

The database queries however fails if fewer than 4 characters have been misidentified due to deletion of the characters in the image processing block, or due to merging of characters. When used for giving specific vehicles access to a barrier area the decision to have an error rate of two characters is in the author's opinion, viewed as acceptable. This is because the likelihood of an unauthorized car having such a similar license plate with all detected characters in their right order is seen as quite small.

3. RESULTS

A set of 108 images, with a resolution of 640×480 pixels is obtained with the camera position set at a distance of between 0.5 and 2 meters from the vehicle. Of the 108 images, 100 images are used to test the developed software, and 68 cases are satisfactorily recognized.

The successful extraction of license plate characters is limited if image has large local variances in pixel ranges in regions other than the license plate, vehicle has bent or warped license plates, characters on the license plate are faded, or had very poor contrast relative to bright regions in the image, or if the vehicle in the image is too close to the camera. Faded plates and texture conflicts present the most difficult test cases. Of the 108 images, 7 images have faded plates, 4 images are obtained with vehicle too close to the camera, 3 images had significantly warped plates. Errors in recognition are largely caused by misclassification of the characters by the template matching algorithm. Efforts to improve its cognitive rate, by use of character standard deviations or Euler numbers to verify the recognized characters, are hindered by an intolerable increase in processing time. The recognition performances for simple visible plates is higher, at about 78%, *i.e.* 62 correct identifications in a sample of 80 images. Contention caused by misclassified digits is resolved by listing all probable license plates from the database.

Fig. 1 shows input image (left) and the resulting extracted license plate characters. The horizontal noise objects violate orientation requirements, and are thus not considered.

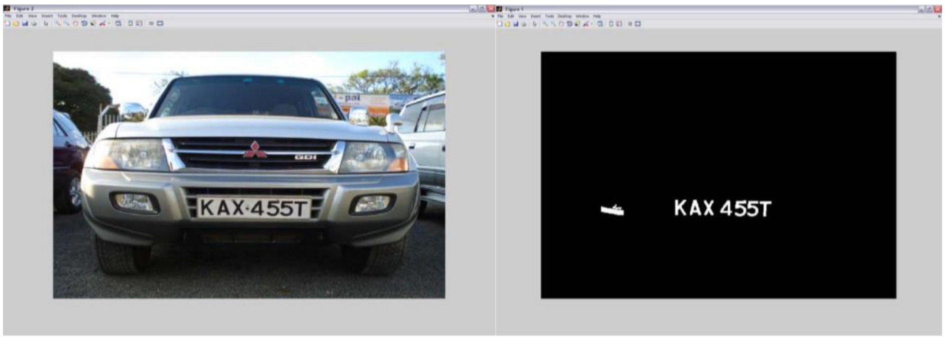


Fig. 1. Input image and the resulting extracted plate characters

Fig. 2 shows input image having large texture regions in areas other than the license plate, due to shadows on car windshield. In this case, extraction of license plate characters is satisfactory.

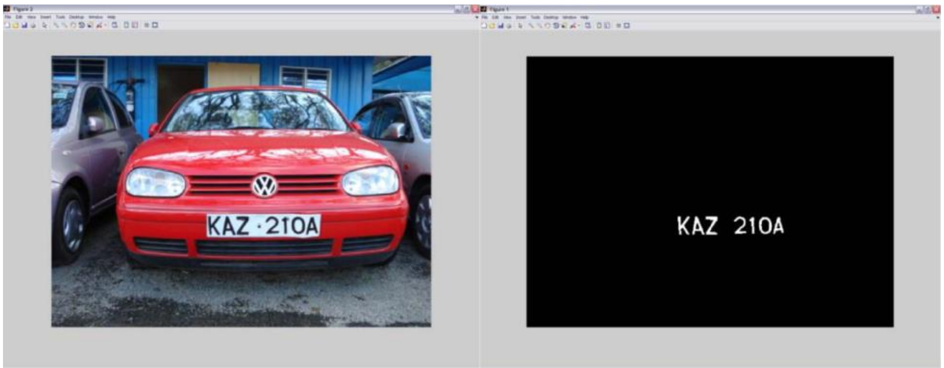


Fig. 2. Input image having large texture regions

Fig. 3 represents the limiting case, showing car on coarse stone background. In this case the texture mask is overwhelmed by the numerous and large changes in local pixel intensities.

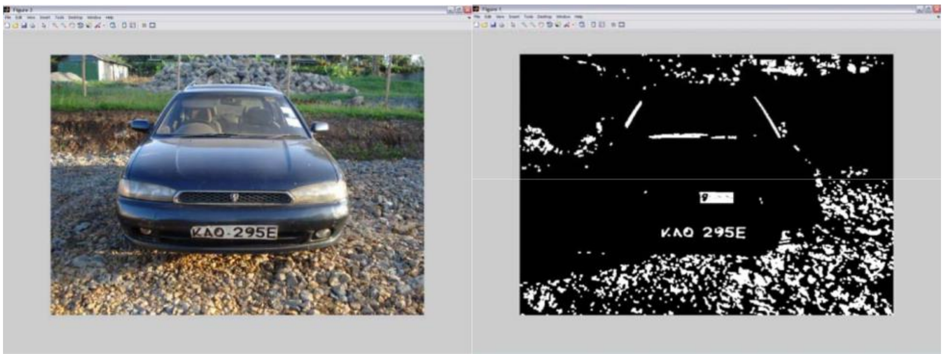


Fig. 3. Car on coarse stone background

Fig. 4 shows images for same vehicle in subsequent frames. While, noise objects imped proper extraction of license plate characters from the first frame, successful extraction from the subsequent frame is possible as shown in Fig. 5.

Fig. 6 shows image with warped plates occluding license plate characters. In this case, an insufficient number of characters (A and 1) are extracted to uniquely identify the vehicle.

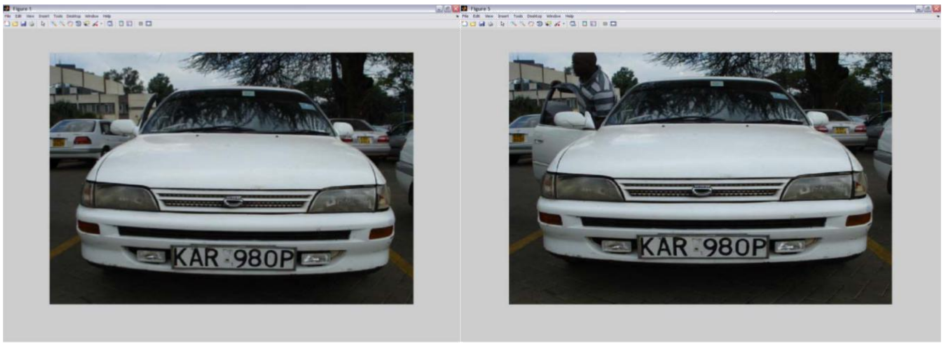


Fig. 4. Images show same vehicle in subsequent frames

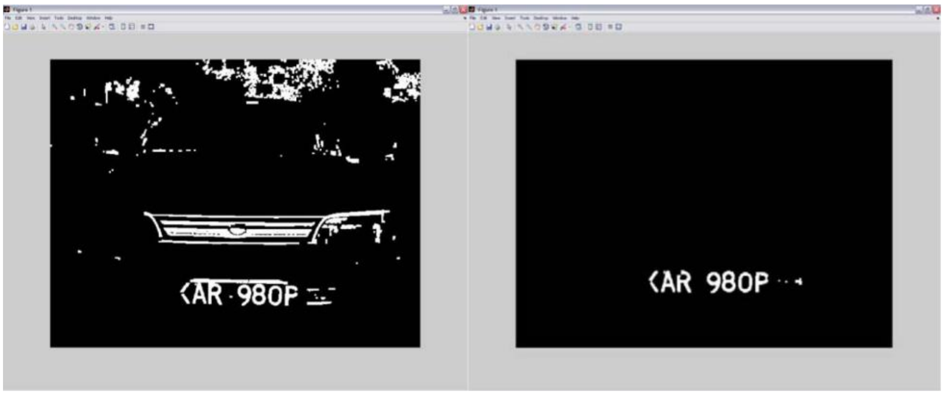


Fig. 5. Extraction of plate characters from subsequent frame

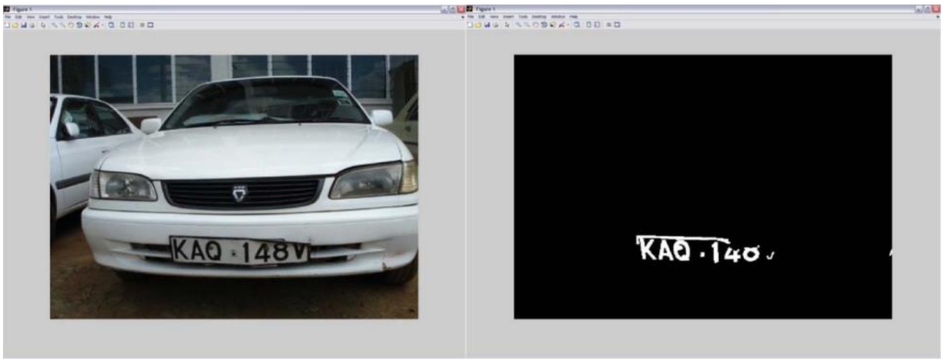


Fig. 6. Image with warped plates occluding plate characters

Fig. 7 shows images of vehicles with faded plates. The extracted license plate characters are too fragmented to be useful as shown in Fig. 8.

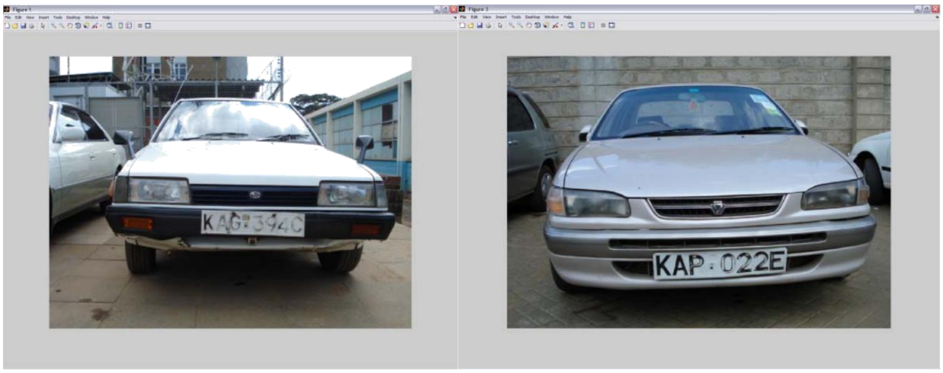


Fig. 7. Images of vehicles with faded plates

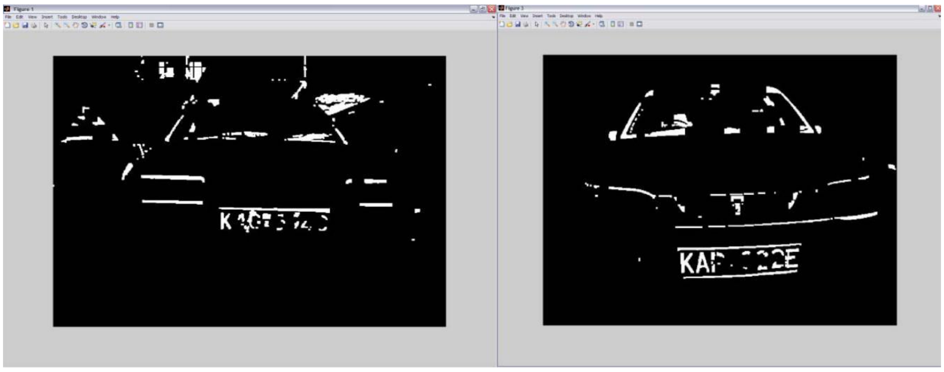


Fig. 8. Extracted license plate characters are too fragmented

Fig. 9 shows input image (left) is too close to the camera. The fixed size of the structuring element used in the top hat filter results in fragmented characters.

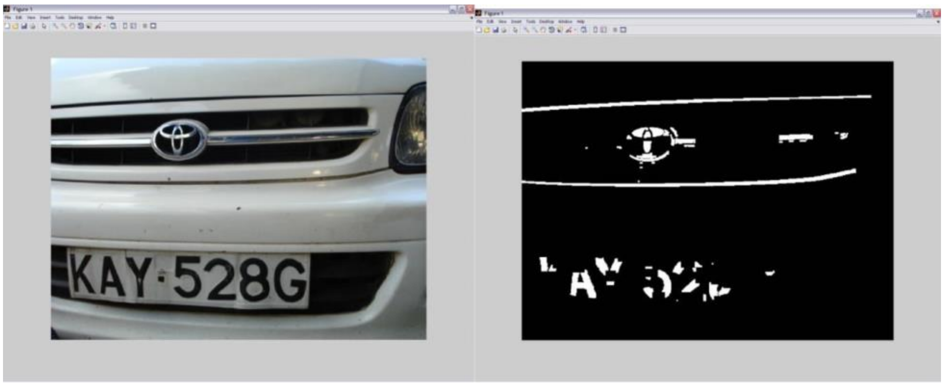


Fig. 9. Input image is too close to the camera

4. CONCLUSION

The car number plate recognition software, comprising of the license plate extraction, character segmentation and optical character recognition modules was designed and implemented. A suitable database with hypothetical user data was also incorporated to complement the system. The ANPR achieved an overall success rate of 68% when tested on 100 of the 108 images, with recognition performances for simple visible plates close to 80%. Results may be improved by refining the recognition stage and testing other classifiers. Different character templates could be used for such refinement of the recognition stage. Future work is intended to be done in improving and testing the system on a larger number of images.

REFERENCES

Badr, A., & Abdelwahab, M. (2011). Automatic number plate recognition system. *Annals of the University of Craiova, Mathematics and Computer Science Series*, 38(1), 62–71.

Lin, Ch.-H., Lin, Y.-S., & Liu, W.-Ch. (2018). An efficient license plate recognition system using convolution neural networks. In *2018 IEEE International Conference on Applied System Invention* (pp. 224–227). Japan: IEEE.

Sharma, G. (2018). Performance Analysis of Vehicle Number Plate Recognition System Using Template Matching Techniques. *Journal of Information Technology & Software Engineering*, 8(2), 1–9.

Silva, S. & Jung, C. (2018). License plate detection and recognition in unconstrained scenarios. In V. Ferrari, M. Hebert, C. Sminchisescu & Y. Weiss (Eds.), *European Conference on Computer Vision* (pp. 593–609). Springer, Cham.

Xie, F., & Zhang, M. (2018). A robust license plate detection and character recognition algorithm based on a combined feature extraction model and BPNN. *Journal of Advanced Transportation*, 2018, 1–14.

Applied Computer Science, vol. 16, no. 1, pp. 85–94
doi:10.23743/acs-2020-07

Submitted: 2020-03-03
Revised: 2020-03-08
Accepted: 2020-03-20

IT company, risk, decision making,
mergers and acquisitions, FTA,ETA

Yehor TATARCHENKO [0000-0002-4036-9318]*,
Volodymyr LYFAR [0000-0002-7860-9663]*,
Halyna TATARCHENKO [0000-0003-4685-0337]*

INFORMATION MODEL OF SYSTEM OF SUPPORT OF DECISION MAKING DURING MANAGEMENT OF IT COMPANIES

Abstract

An information model has been carried out, with the help of which it is possible to implement methods that ensure the growth of competitiveness of IT companies. Growth conditions for companies provide mergers and acquisitions (M&A). The analysis of the data obtained as a result of the P&L financial report is mainly based on current indicators and can be partially used to prolong economic indicators for a certain (most often limited) period. The authors propose using methods for assessing stochastic indicators of IT development processes based on the solution of a number of problems: (1) Development of models to assess the impact of indicators in the analysis of the financial condition of companies; (2) Creation of an information model and methods for processing current stochastic data and assessing the probability of the implementation of negative and positive outcomes.

1. INTRODUCTION

The development of information technology projects includes procedures based on a special project-driven approach to all stages of the project life cycle (Darnall & Preston, 2016). At the same time, such projects are among the most risky investments.

* Volodymyr Dahl East Ukrainian University, Faculty of Information Technology and Electronics, Department of Programming and Mathematics, Tsentralnyi Ave., 59A, Severodonetsk, Luhansk Oblast, Ukraine, 93400, gosahi@gmail.com, lyfarva61@ukr.net, tatarchenkogalina@gmail.com

The current business climate is characterized by continuous competition, profit-shifting and rapidly changing technologies. Moreover, the instability in the market of IT services is often caused by a random cause independent of enterprise management (Pagach & Warr, 2011). The use of mergers and acquisitions (M&A) to manage the market for services and production in the field of information technology can significantly stabilize the risks of financial investments in the IT market (What are the Main Valuation Methods?, 2019). One of the important components of mergers and acquisitions is a qualitative assessment of the value and condition of the company associated with the development of IT. The main objective of mergers and acquisitions is to increase the value of objects created by the company. At the same time, the cost of a business combination should be greater than the total values of the merger component. Violation of this rule is possible if the risk assessment of the state of companies is incorrect. Of course, first of all, from the indicators of the "value" of the company, it is necessary to determine its financial value and solvency. In this case, it is necessary to analyze simultaneously indicators that affect the financial condition. It is necessary to try to achieve the highest possible synergies (additional value resulting from M&A).

The solution of these problems is associated with the use of a more complex methodological database of data processing, a wide range of variables, the use of various models and algorithms for assessing values and costs (What are the Main Valuation Methods?, 2019).

However, even in this case, obtaining reliable estimates in the analysis of the data, provided that there are a large number of different factors, may be questioned in connection with the stochastic nature of the input data. It is in connection with the above that a number of important scientific and technical problems can be distinguished that can be solved by combined methods of mergers and acquisitions and methods of assessing the probability of occurrence of consequences from mergers and acquisitions.

2. GENERAL REPRESENTATION OF MODELS AND METHODS FOR ASSESSING THE STATUS OF IT COMPANIES

To solve the problems of assessing the status of IT companies, a simulation-stage modeling method is proposed, based on the sequential formalization of logical causal relationships of events that may be present in the structural model of the associated scenario development processes when introducing organizational, economic, technical and other solutions during the implementation of M&A.

Processes of mergers and acquisitions give rise to many technological, organizational, economic, psychological, informational events that can lead to both positive and negative consequences. The main task of modeling is to select such solutions, leading to an integrated positive effect. For effective modeling, it is proposed to create a set of models that meet the following requirements:

- normative – from the reference (description of the class of the object) to a specific object;
- dynamic (imitation);
- material and procedural;
- stochastic and substantial.

When solving the problems of simulation-step-by-step modeling, M&A processes are combined with methods of rationalizing company resources according to their importance (economic effect). The authors propose the introduction of a hybridization of ARIS (Architecture of Integrated Information System) and structural modeling methods of the IDEF class and the principles of ABC analysis (Ultsch & Lötsch, 2015; Kringel et al., 2017; Iovanella, 2017; Pawelek, Pociecha & Baryla, 2017).

Structural methodologies are represented by the following models:

- Function Modeling – functional modeling using graphical tokens IDEF0 combining a set of interconnected functions (blocks). Typically, IDEF0 is used in the first step in the analysis of any structured system. This method is the next stage in the development of the well-known language for the description of functional systems SADT (Structured Analysis and Design Technique) (Draft Federal Information Processing Standards Publication 183, 1993);
- Information Modeling – IDEF1 modeling of information flows inside systems, allowing to display and analyze their structure and mutual relations. IDEF1 Extended – Data Modeling – database modeling methods based on the entity-relationship model. The IDEF1 method allows you to build a structural data model equivalent to the relational model in the 3rd normal form. IDEF1X diagrams are used by many CASE tools (in particular, ERwin, Design/IDEF) (Draft Federal Information Processing Standards Publication 184, 1993).
- Process Description Capture (documentation of technological processes) – methods for documenting processes that occur in a system or projects that describe the scenarios and logical sequence of operations for each important process or event. IDEF3 has a direct investigation sequence associated with IDEF0 so that each function can be decoded in the form (protocol) of a separate IDEF3 process;
- Object-Oriented Design – the methodology for building object-oriented systems (IDEF4) is proposed to be replaced by the methods of “fault trees” (FTA) (IEC 60300-3-9:1995; SS-IEC 1025:1990);
- Ontology Description Capture – It is proposed to replace the standard of ontological research of complex systems (IDEF5 methodology) with the methods of “event trees” (ETA).

In parallel with structural modeling methods, you can use ABC analysis.

For each stage of the life cycle of companies, the development and combination of structural models of different levels and purposes is carried out. Their combination into one logical form makes it possible to create a digraph of the state of companies on the convolution / development of chains of the graph to the level of tree branches.

The general approach to simulation-stage modeling shown in Fig. 1.

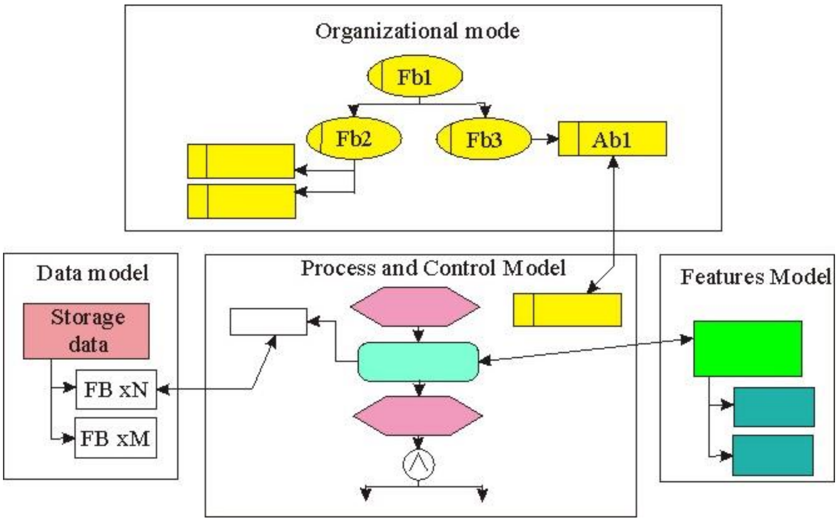


Fig. 1. Synthesis of the simulation-stage model

The following sequence is formed into a common integrated model: Organization chart (Organizational chat) ‖→ Function Tree ‖→ process / control models (diagram eEPC (extended Event driven Process Chain) → IDEF0 → IDEF1X → IDEF3 → FTA → ETA. In parallel, ABC analysis methods are being implemented to determine the financial consequences of structural events.

In the described structures:

– the IDEF0 model consists of sets of hierarchical linked diagrams. In the diagram, the blocks are combined by arcs (a subgraph of the functional model) and the output arcs of some blocks can be inputs of others. Arcs with one free end have a coil or a receiver outside the diagram. To determine the external arcs use the following notation: I (Input), C (Control), O (Output) M (Mechanism);

– the data is represented by a model of the IDEF1 Extended class in such a way that all input data and output variables must be normalized and fully defined in other models;

– IDEF3 – models are used to document technological (information) processes where it is important to take into account the sequence and logical direction of the process.

The development of a simulation-stage model (SSM) of the studied companies, in contrast to the existing methods of discrete-event modeling, allows one to take into account causal relationships between events and processes that are stochastic. The simulation-stage model allows you to analyze the probability of the implementation of events and processes, and also, thanks to the step-by-step examination of these processes, take into account the dynamic characteristics.

First, based on the analysis of event and process structures, experts formalize and link cause-effect chains of events of the SSM model in the FTA format, using the top-down analysis method. The number of "failure trees" reflects the complete set of events that have consequences ("upper events", for example: implementation of development at a given time).

The top event can be either a specific achievement with a positive effect, or the failure of this achievement. The difference in this case is the probability of failure P_f, or the probability of success $Ps = 1 - P_f$.

For each "fault tree" reaching the "upper event", one can concatenate this event as the initial condition of the "event tree". Starting with the initiating event, the binary branching of the "tree of events" is built, which reflects the logic of the development of various scenarios, taking into account stochastic indicators and the effectiveness of the means of influence (decisions). In this case, scenario analysis branches are formed by an upward analysis, which result in positive or negative consequences.

Integral indicators of risk that are created by various scenarios include:

1. The probability of negative development scenarios (total n), which are due to j-processes of i-solutions:

$$P_t^n = \sum_i \sum_j P_{ij}^n \tag{1}$$

2. The probability of a positive development scenarios (total p):

$$P_t^p = 1 - P_t^n \tag{2}$$

3. Expected total loss:

$$ED = \sum_i \sum_j P_{ij}^n \cdot Us_{ij} \tag{3}$$

   where: P_{ij}^n – the probability of implementing the j-th negative scenario that occurs when making the i-th decision; Us_{ij} – loss from the j-th process of the i-th solution.

4. Expected total profit:

$$EP = \sum_i \sum_j P_{ij}^p \cdot In_{ij} \tag{4}$$

where: P_{ij}^{p} – probability of implementation of the j-th positive scenario, which is realized when the i-th decision is made; In_{ij} – profit from the j-th process of the i-th solution.

To determine an acceptable level of risk, an overall expected effect is proposed. That is, if the sum of the expected damage and the expected profit is positive, then consider that the overall risk as such does not pose a threat. However, if the expected total income is less than that obtained before the introduction of mergers and acquisitions, this confirms the assumption that such procedures do not make sense (more harmful).

An assessment of the risk of financial consequences is determined on the basis of an analysis of their likelihood, as well as anticipatory risk actions in the company and the restoration of positive processes.

3. INFORMATION MODEL OF EVALUATION OF THE PROBABILITY OF EVENT DEVELOPMENT IN IT COMPANIES

Support for solutions that are optimal in the sense of Pareto is based on the likelihood of implementing scenarios for the development of IT companies in situations of conflict and uncertainty. By "conflict" is meant the competitive development of various indicators (positive and negative, profit and loss) of the current state of the analyzed objects.

In this case, the risk is considered as the occurrence of certain events with a certain probability. Moreover, all events can be quantitative (for example, cost) or qualitative (for example, permissible, unacceptable) indicators.

Using the concepts of game theory, the state of the analyzed objects was presented as a matrix of possible states obtained on the basis of event trees. Each strategy Sj, represented by a proposed or predictable set of impacts on an object, is evaluated either quantitatively through profitability or loss indicators, or qualitatively by characteristic indicators of levels of positive or negative consequences. At the same time, mixing ratings is unacceptable.

To determine the quantitative values of the probability of initiating events that determine the initial state of the subsystem, FTA analysis methods were implemented, and to determine the development of scenarios of probable results, it was used in the upward ETA analysis.

The above is implemented by formatting information and data conversion processes in such a way that:

- to obtain criteria parameters of limiters of risk indicators;
- evaluate current indicators of risk due to processes within the companies under study;

- establish a correspondence between risk indicators and input events of influence on the state of companies and state changes in connection with disturbances;
- determine the set of scenarios of events taking into account the probability of occurrence of x conditions;
- determine the impact of events that constitute a negative scenario;
- analyze the cause-and-effect processes of the emergence and development of scenarios and identify the many solutions leading to this, for the analysis of solutions that can increase the positive effect of mergers and acquisitions;
- perform the search process for Pareto optimal solutions based on risk indicators and economic consequences, and determine the set of optimal solutions.

In structural step-by-step modeling, chains of a state graph are formed and separated that simulate scenarios of the occurrence and development of events, for which it is possible to determine the consequences taking into account the probability, which is mathematical modeling of stochastic processes.

Each chain that is defined is directional, connected, fully defined, eulerian and allows you to get all the risk indicators. The sets of chains intersect with the subsets of states and form parts of the graph indicated in formula 5.

The synthesis of the simulation-stage model is implemented by connecting the input and output parameters at the nodes of the chain of cause-effect relationships. The harmonization of information presentation formats is ensured through a structured presentation of data presented in xml format.

The synthesis mechanism proposed in the work allows combining the logical modeling of events and processes and the modeling and analysis of economic indicators into a single information technology. This technology, unlike the existing ones, allows you to use a simulation-stage model of the state of companies and introduce consistent calculations to determine the integrated risk indicators, and analyze these events to determine and comparative analysis of the consequences.

4. MATHEMATICAL MODEL OF PRACTICAL EVALUATION OF COMPANIES

The generalized mathematical model of the yaw assessment is based on the specific processing of data represented by a tuple:

$$MTR = \langle Ep, R.Inv, M_n, M_p \rangle \tag{5}$$

where: $Ep = \{ep_j\}$ – many events occurring in companies; $R = \langle \vartheta, P, D \rangle$ – tuple of risk specific to the processes under consideration; $Inv \subseteq W(O) \times P$ – correspondence between input events and the probability of transition to different states by impact; $M_p = \{mp_z\}$ – many positive consequences of the development of events $z \in 1 \dots A$ in monetary terms; $M_n = \{ma_c\}$ – many negative consequences in monetary terms.

The task is formalized as follows:

The risk function has the form:

$$R = \langle \vartheta, P, D \rangle \tag{6}$$

where: ϑ – many influences that defines scenarios; $P = \left[P_n, P_p\right]$ – set of probabilities of possible consequences (negative and positive);

Let n companies that have *i* states be considered, then for any *i*-th state the risk D of consequences is determined: $R_i = \langle \vartheta, P, D \rangle_i$.

Believed to be known:

– deterministic models of the development of processes that can lead to the *i*-th state:

$$Fne_{ij}\colon \overrightarrow{S_{\imath\jmath}} \to \overrightarrow{\Phi_{\imath\jmath}}, j = 1 \dots J \tag{7}$$

where: j – (a set of elementary events leading to certain states), $\overrightarrow{S_{\imath\jmath}}$ – vector of parameters that defines the initial state for the j-th event; $\overrightarrow{\Phi_{\imath\jmath}}$ – vector of phase variables of elementary processes in the system that may occur in the *i*-th state;

– model for assessing the probability of the development of stochastic elementary events: $Pa_{ij}\colon \left(\vec{S}, \vec{\Phi}\right)_{ij} \to \overrightarrow{P_{\imath\jmath}}$, $j = 1 \dots J$, where $\overrightarrow{P_{\imath\jmath}} = \left[P_{ij}^{p}, P_{ij}^{n}\right]$ – vector of probabilities of positive and negative consequences.

The model of determining the influence of events on the state of the system from decisions made to analyze and predict economic consequences containing:

– model for assessing the probability of occurrence of rare events in the *i*-th system in the form FTA (total *k* trees)

$$\gamma_k\colon \left(\{\overrightarrow{P_{\imath\jmath}}\}, \overrightarrow{P_{k\imath}}\right) \to \overrightarrow{P_{k\imath}}, \tag{8}$$

– model simulating the development of events in the form ETA

$$\mu_k\colon \left\{\left(S, \Phi, \overrightarrow{P_k}\right)_i, \overrightarrow{\vartheta_k}\right\} \to M_{ki} \tag{9}$$

where: $S_i = \{\overrightarrow{S_{\imath\jmath}}\}$, $\Phi_i = \{\overrightarrow{\Phi_{\imath\jmath}}\}$, M_{ki} – integral indicators of profit from the *k*-th state of the scenario.

The total set of FTA and ETA associations for all i-subsystems of the companies under investigation, as well as indicators of expected loss and profit can be represented by a generalized graph MTR. The graph is subject to analysis and processing of indicators to search for branches of scenarios and assess their consequences. This is the basis for a comparative analysis when making decisions optimized in the Pareto sense for multicriteria indicators.

Models of the final scenarios of accident development are based on mathematical modeling of sequential processes and events and are contained in the state graph by setting up a serial connection of input and output events, which are determined by experts. Moreover, the graph chains have weight indicators that reflect the level of stochasticity of certain events. It should be noted that the different goals of the processes in multi-parameter estimates are directed in different directions. In this regard, it is necessary to apply methods of distributing the importance of various sinks of events for the benefits that are established by experts and make up the meaning of Pareto optimization as a sequence of dominant decisions.

The task for optimization according to many criteria is considered as an optimization problem at the same time for all isolated criteria. Searching for a set of solutions $\vec{x} \in X$, such that are minimized by all these criteria in a sense. That is, we consider a sequential optimization problem corresponding to the conditions: $g^{(k)}(x) \rightarrow min, k = \overline{1, N}$, on condition $x \in X$. In this case, the criteria $g^{(k)}(x)$ there are *partial criteria*. Their sets can be considered "vector criteria" $G(\vec{x}) = \left(g^{(1)}(\vec{x}), \dots, g^{(N)}(\vec{x})\right)$ which are subject to optimization for the benefits of components established by experts.

4. CONCLUSIONS

As a result of the research, a mathematical information model and methods for analyzing the stochastic and determinate components of the risk of consequences that affect events on the development of the life cycle of the development of IT companies were developed and agreed. These methods are based on the construction of a directed graph with sequences of logical cause-effect relationships of initial and subsequent events and influences.

The mathematical model is proposed for processing information flows that reflect the state of IT companies. The model takes into account the probability of the development of positive or negative consequences of decisions. The novelty of this approach lies in the proposed mechanism for the joint use of FTA and ETA, as well as a comparative analysis of the expected extent of the consequences. Unlike existing information models for analyzing the state of IT companies, this model takes into account stochastic characteristics of processes that have a significant impact on the consequences of decisions regarding mergers and acquisitions.

REFERENCES

Darnall, R., & Preston, J. M. (2016). *Project Management from Simple to Complex*. University of Minnesota Libraries Publishing.

Draft Federal Information Processing Standards Publication 183. (1993). *Integration Definition For Function Modeling (IDEF0).*

Draft Federal Information Processing Standards Publication 184. (1993*). Integration Definition For Information Modeling (IDEF1X).*

IEC 60300-3-9:1995. (1995). *Dependability management – Part 3: Application guide – Section 9: Risk analysis of technological systems*.

Iovanella, A. (2017). Vital few e trivial many. In *Il Punto* (pp. 10–13).

Kringel, D., Ultsch, A., Zimmermann, M., Jansen, J. P., Ilias, W., Freynhagen, R., & Resch, E. (2017). Emergent biomarker derived from next-generation sequencing to identify pain patients requiring uncommonly high opioid doses. *The pharmacogenomics Journal*, *17*(5), 419–426, doi:10.1038/tpj.2016.28.

Pagach, D., & Warr, R. (2011). The Characteristics of Firms That Hire Chief Risk Officers. *The Journal of Risk and Insurance*, *78*(1), 185–211.

Pawelek, B., Pociecha, J., & Baryla, M. (2017). ABC Analysis in Corporate Bankruptcy Prediction. In *Abstracts of the IFCS Conference* (p. 17). Tokyo, Japan.

SS-IEC 1025:1990. (1990). *Fault tree analysis (FTA).*

Ultsch, A., & Lötsch, J. (2015). Computed ABC analysis for rational selection of most informative variables in multivariate data. *PLOS One*, *10*(6), e0129767. doi:10.1371/journal.pone.0129767

What are the Main Valuation Methods? (2019). Retrieved August 12, 2019 from https://corporatefinanceinstitute.com/resources/knowledge/valuation/valuation-methods

Applied Computer Science, vol. 16, no. 1, pp. 95–102
doi:10.23743/acs-2020-08

Submitted: 2020-03-14
Revised: 2020-03-20
Accepted: 2020-03-25

remote health monitoring, fall detection, design

Mohanad ABDULHAMID[*], *Deng PETER*[**]

REMOTE HEALTH MONITORING: FALL DETECTION

Abstract

Falling is a serious health issue among the elderly population; it can result in critical injuries like hip fractures. Immobilization caused by injury or unconsciousness means that the victim cannot summon help themselves. With elderly who live alone, not being found for hours after a fall is quite common and drastically increases the significance of fall-induced injuries. With an aging Baby Boomer population, the incidence of falls will only rise in the next few decades. The objective of this paper is to design and create a fall detection system. The system consists of a monitoring device that links wirelessly with a laptop. The device is able to accurately distinguish between fall and non-fall.

1. INTRODUCTION

Healthcare systems in the world have undergone tremendous evolution in the last 50 years. In the early 1960s, we had computers in the form of mainframes being incorporated into healthcare systems. However, there were some problems met from their usage. These mainframes were very few, expensive, large in size and consumed a lot of electrical power and as a result, they had to be shared by several hospitals since independent ownership wasn't feasible an idea to be considered at all (Huang & Newman, 2012).

Come 1970s to early 1990s, there were enormous changes in terms of size and cost for computers & some of the hospital equipment that were invented and innovated at the time. Hospitals were thus able to acquire smaller sized computers and be able to easily operate their independent healthcare systems. Also, in this period, we had the invention of the internet which had a positive impact on health systems. Communication amongst hospitals, their staff and patients greatly improved.

[*] Al-Hikma University, Karada Kharidge, Baghdad, Iraq, moh1hamid@yahoo.com
[**] University of Nairobi, P.O. Box 30197-00100, Nairobi, Kenya, researcher12018@yahoo.com

Hospitals were also able to upload and store some of their data (especially patients' data) online so as to make their accessibility easy for authorized personnel (personnel need not be physically present at the premise since all they needed was just a computer, internet connection & and the necessary access password) (Gong, Wang, Zhang & Wang, 2017).

Despite all of the tremendous changes, there was still the issue of affordability of the treatments offered by the hospitals (that were properly equipped with computers & hospital equipment) from the patients' point of view. This had a negative impact on both the hospitals & patients. Hospitals were not getting that large enough a number of patients for treatment while patients were opting for alternatively cheaper treatment plans (which were not that good enough compared to that of properly equipped hospitals) (Saranya, Preethi, Rupasri & Veena, 2018).

Then came the mid-1990s to early-2000 and present where significant technological advancements have taken place. This has seen to great improvements in the healthcare systems with the diversification of remote health monitoring which by definition, is a form of technology which allows a patient to use a mobile medical device to perform tests from outside a clinic and collects the medical and health data to securely transmit to healthcare professionals for remote assessment. As a result, patients, especially the chronically ill, elderly or disabled are able to have increased healthcare access at their homes with decreased healthcare delivery costs (Malasinghe, Ramzan & Dahal, 2019).

The scope of this paper is limited to fall detection and remote viewing of the collected data. Fall detection algorithm is developed and implemented through programming on Arduino Uno board. Remote viewing of the data is done on a laptop after having fall detection data wirelessly transmitted to it.

2. Design and methodology

2.1. The algorithm design

To detect fall along an axis, the acceleration magnitude is considered. This is achieved by a magnitude vector. Consider:

$$AM = \sqrt{{a_x}^2 + {a_y}^2 + {a_z}^2} \tag{1}$$

where AM is acceleration magnitude.

With the accelerometer output data, the angle change can also be calculated using the dot product. To achieve this, the instantaneous vector and a reference vector are introduced. Instantaneous vector is given by

$$a = (a_x, a_y, a_z) \tag{2}$$

Reference vector is generated when a user stands up. It is given by:

$$b = (b_x, b_y, b_z) \tag{3}$$

Using both the instantaneous vector and reference vector in the following formula:

$$a \cdot b = |a| \cdot |b| \cdot cos\theta \tag{4}$$

Making the angle as subject

$$\theta = cos^{-1}\left(\frac{a \cdot b}{|a| \cdot |b|}\right) \tag{5}$$

In the event of falling, one experiences a momentary free-fall then a large spike in acceleration. In the flow-chart shown in Fig. 1, we have two decision figures: lower threshold AM and upper threshold AM. The algorithm runs in the following manner. First, it checks whether the lower threshold value has been broken by the AM and if so, it then quickly checks whether the upper threshold is broken within a span of 0.5s. If it's not broken, we go back to data collection and if so, the algorithm recognizes this event as a fall. This algorithm's strength is that it requires two AM thresholds to be broken by an activity for a fall to occur.

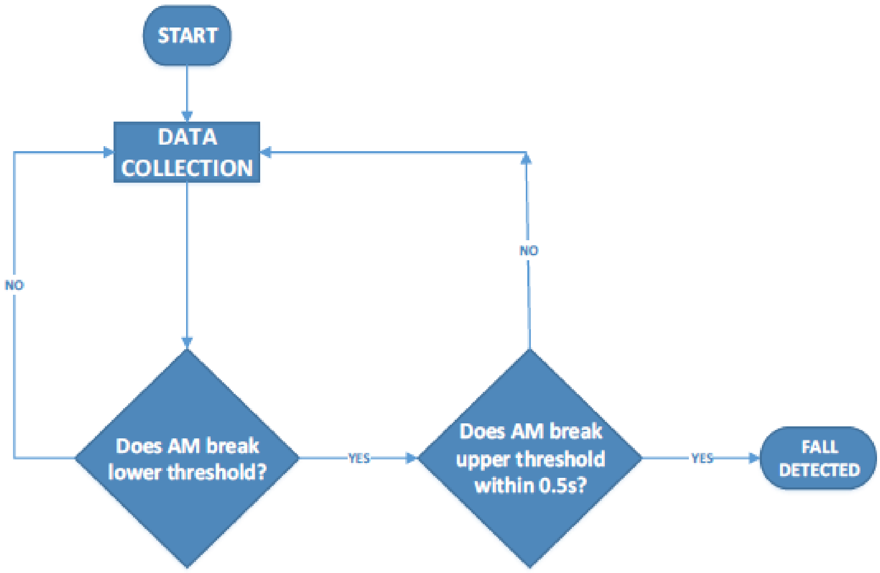

START
DATA COLLECTION
NO
NO
Does AM break lower threshold?
YES
Does AM break upper threshold within 0.5s?
YES
FALL DETECTED

Fig. 1. Fall detection algorithm

2.2. Components

2.2.1. Arduino Uno R3

The Arduino Uno is a very popular board among hobbyists and is the microcontroller board of choice when building small model projects. Because of this, there are extensive tutorials and open source examples available to facilitate learning and familiarizing oneself with the board. In addition to this, we choose this board because of the following characteristics:

1. Operating voltage – The operating voltage of 5V with a 3.3V option is appropriate because both our sensor boards and Bluetooth module operate under 5 or 3.3V power and output readings in the range of 0–5V.
2. Input voltage – The board has a built-in voltage regulator that allows an input voltage range of 7–12V, which is suitable because we plan to power the board with a 9V battery.
3. Memory – The flash memory (32KB) is appropriate because our algorithm programs can be fairly long and require a decent amount of memory on the microcontroller to store them. The Static Random Access Memory(SRAM) (2KB) is a little on the low side, but the algorithms can work around this by not storing too many variables, so as to not exhaust the SRAM capacity.
4. Specialty pins – The Arduino Uno comes with RX/TX pins, which will be used for serial communication with our Bluetooth module. The board also comes with I2C compatible pins, which will be crucial to interface with our digital accelerometer.

2.2.2. Bluetooth module

For wireless data transmission, we choose the HC-05 module which is an easy to use Bluetooth SPP (Serial Port Protocol) module, designed for transparent wireless serial connection setup. The HC-05 Bluetooth module can be used in a Master or Slave configuration, making it a great solution for wireless communication. We choose this module namely because its pins and power are 5V compatible. It also supports RX/TX serial communication from 9600 to 115200bps (bits per second, baud rate), which makes it fully compatible with our Arduino Uno R3 board.

2.2.3. Sensor

For appropriate fall detection, we choose an accelerometer**.** The model selected is the ADXL345 triple axis digital accelerometer. It has a wide G range (up to ±16g). The range is very wide considering some severe falls are rated at 8 g's. Since it is a digital sensor, the resolution can be adjusted and there is less voltage noise, and less calibration. The ADXL345 gives tri-axial data and requires a minimum of 3.3V power, is I2C compatible and thus our microcontroller board can interface with it correctly.

2. 3. Overall system design

2.3.1. Circuit setup

In order to achieve the assumed goal an appropriate electronic circuit was developed. The overall circuit setup is shown in Fig. 2.

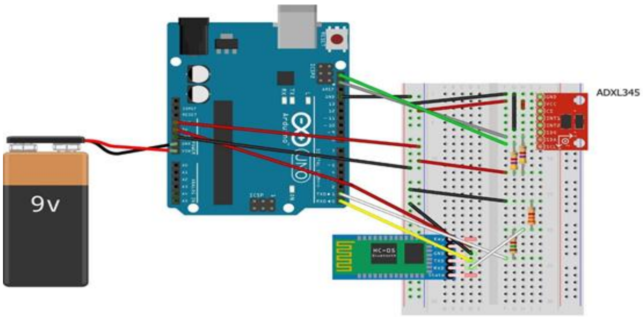


Fig. 2. Overall circuit setup

2.3.2. Block diagram

In the proposed solution, we have the collection of data by the accelerometer, processing of the same by the micro-controller and ascertaining of whether a fall has occurred (Fig. 3 – section A). Our algorithm, is run by the micro-controller. The Bluetooth module receives fall detection data from the micro-controller and transmits the same to a laptop computer where we realize remote viewing of the data. The overall block diagram is shown in Fig.3.

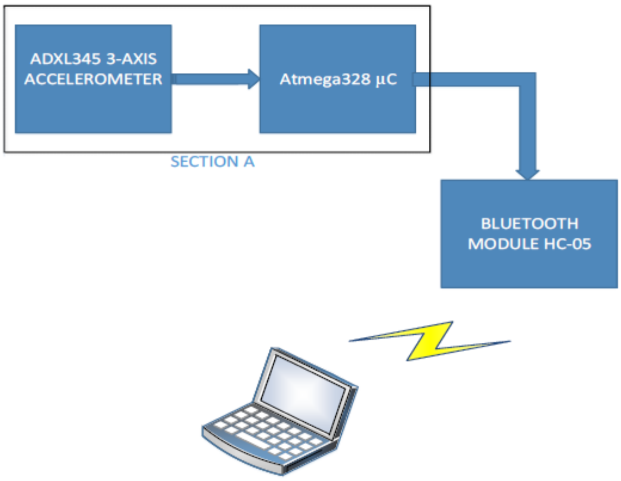


Fig. 3. Overall block diagram

3. RESULTS

The graphs below represent the fall and non-fall data that we collect during our testing phase. They present a comparison between the kind of graphs we expect to see in a fall and non-fall scenarios. From our algorithm, we set the upper threshold value to 2g's so as to ensure that all acceleration values from fall activities break it just as illustrated in Fig. 4. Also, we see that the set lower threshold value of 0.4 g's has been broken. In Fig. 5 and Fig. 6, we have graphs for non-fall activities. In both cases, we see that their acceleration magnitude does not exceed the set upper threshold value of 2g's.

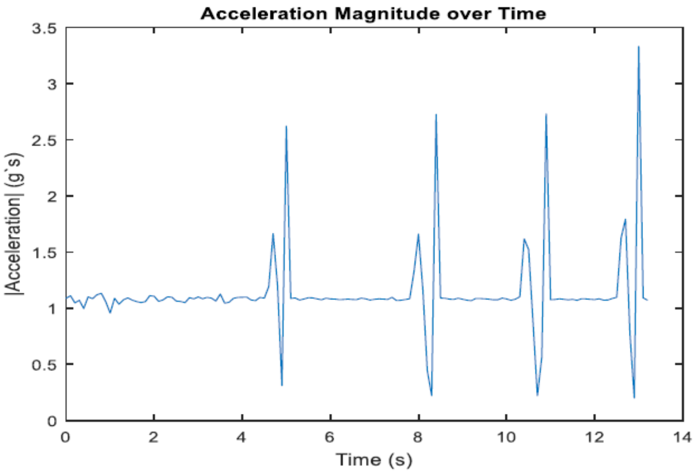


Fig. 4. Graph of falling down

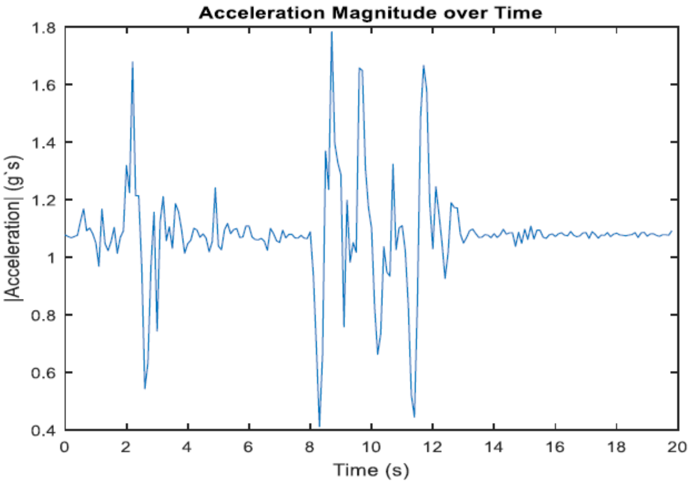


Fig. 5. Standing up and sitting down hard

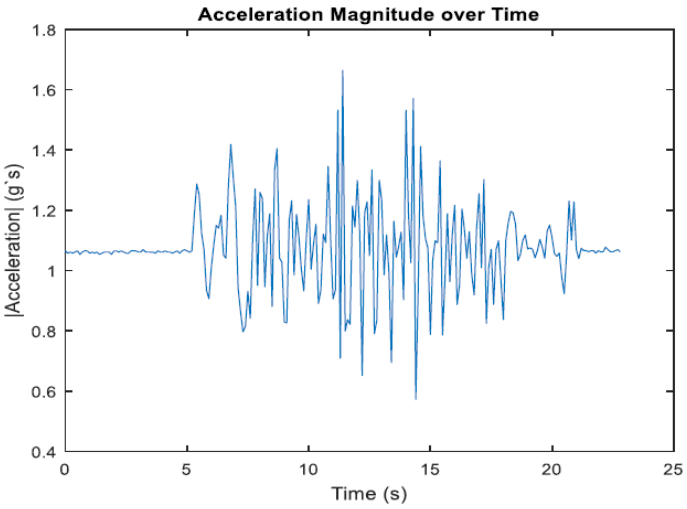


Fig. 6. Walking

4. CONCLUSION

The objective of this paper was to design a fall detection system that links wirelessly with a laptop computer (where we have remote viewing of the collected data). By the conclusion of this paper, we had achieved the primary goal of creating a working system able to recognize falls from non-falls, while wirelessly synched with a laptop.

With this paper, there are some areas for future development. From the commercial point of view, improvements would include: having the system housed in a proper and well-designed casing to prevent its damage in the occurrence of a fall, establishing emergency contacts though the PC-side by sending text messages, reduction of the size of the system by using custom printed circuit boards and Lithium-ion batteries, and porting the PC-side programming onto a mobile phone to realize complete mobile communication. Also, we could add a gyroscope and Global Positioning System(GPS) module to the setup. Both would greatly improve on the overall efficiency of the system

REFERENCES

Gong, S., Wang, Y., Zhang, M., & Wang, C. (2017). Design of remote elderly health monitoring system based on MEMS sensors. In *2017 IEEE International Conference on Information and Automation (ICIA)* (pp. 494–498). Macau: IEEE.

Huang, Y., & Newman, K. (2012). Improve quality of care with remote activity and fall detection using ultrasonic sensors. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 5854–5857). San Diego, CA: IEEE.

Malasinghe, L., Ramzan, N., & Dahal, K. (2019). Remote patient monitoring: a comprehensive study. *Journal of Ambient Intelligence and Humanized Computing*, 10(1), 57–76.

Saranya, M., Preethi, R., Rupasri, M., & Veena, S. (2018). A survey on health monitoring system by using IOT. *International Journal for Research in Applied Science & Engineering Technology*, 6(III), 778–782.