# Gender Prediction of Journalists from Writing Style

Peshawa J. Muhammad Ali[1], Nigar M. Shafiq Surameery[1], Abdul-Rahman Mawlood Yunis[2] and Ladeh Sardar Abdulrahman[1]

[1]Software Engineering Department, Koya University
Daniel Mitterrand Boulevard, Koya KOY45 AB64, Kurdistan Region - Iraq

[2]Canada Revenue Agency, Ottawa, Ontario, Canada

***Abstract*—Web-based Kurdish media have seen a tangible growth in the last few years. There are many factors that have contributed into this rapid growth. These include an easy access to the internet connection, the low price of electronic gadgets and pervasive usage of social networking. The swift development of the Kurdish web-based media imposes new challenges that need to be addressed. For example, a newspaper article published online possesses properties such as author name, gender, age, and nationality among others. Determining one or more of these properties, when ambiguity arises, using computers is an important open research area. In this study the journalist's gender in web-based Kurdish media determined using computational linguistic and text mining techniques. 75 web-based Kurdish articles used to train artificial model designed to determine the gender of journalists in web-based Kurdish media. Articles were downloaded from four different well known web-based Kurdish newspapers. 61 features were extracted from each article; these features are distinct in discriminating between genders. The Multi Layer Perceptron (MLP) artificial neural network is used as a classification technique and the accuracy received were 76%.**

***Index Terms*—Gender identification, Kurdish media, Neural networks, Text mining.**

## I. INTRODUCTION

Telecommunication and Internet sectors have witnessed a rapid development in Iraq after year 2003. According to Internet World Stats (www.internetworldstats.com), the number of Internet users is estimated to be 2.5 million users in 2012 while this number was 1.3 million in December 2011. This growth in the telecommunication and Internet sectors is mainly due to the vast growth of these sectors in Iraqi Kurdistan region. According to the Kurdistan Region's Ministry of Communication and Transportation web site (www.moc-krg.com), there are two major Internet companies along with 21 other smaller companies operate in Kurdistan region providing Internet services.

Alongside the rapid growth of the Internet infrastructure, a large number of Kurdish websites and web-based newspapers have appeared, as well. Some of these newspapers have become very well-known and they have a large user base. Furthermore, the growth in Internet and telecommunication sectors has led to easy access, low in pricing and enabled pervasive usage of social networking. These changes, in turn, made it possible for anyone to publish and share opinion. Nowadays, not only main stream media can publish articles but any person can do so using Facebook pages, blogs or a personal web page. With the easy access, inexpensive electronic gadgets, and pervasive usage of social networking comes the problem of anonymous publishing and identity faking problem. Currently, fake personalities on social networks like Facebook and Tweeter are wide spread phenomenon.

*Gender identification* research falls under a wider research area known by *authorship identification.* Studies in the latter field include the attribution of disputed Shakespearian poetries done by (Efron and Thisted, 1976), and (Merriam, 1996). On the other hand, early works on gender identification, which is about an examination of a specific part of authorship identification, has been conducted by (Lakoff, 1973), and (Labov, 1990). These works have concluded that there are differences in the male and female writing style.

The objective of this study is to identify authors' gender from their online writing style. More precisely, the main question of this research work is that "Would be possible to distinguish male writers from female writers based on their online writing?" This question has already been answered for languages such as English and Arabic. This paper applied a technique not used before to a language that never examined for this purpose. To the best of our knowledge, this effort is the first of his kind. Thus, this paper has opened a new research direction in this field for Kurdish language.

ARO, The Scientific Journal of Koya University
Volume I, No 1(2013), Article ID: ARO.10031, 7 pages
DOI:
Regular research paper
Received 26 September 2013; Accepted 12 November 2013
Corresponding author's e-mail: peshawa.jammal@koyauniversity.org

The rest of this study is organized as follow: in Section II the most relevant work to this study are reviewed. In Section III, study challenges and assumptions are discussed. In Section IV, the solutions and experimental results are presented, and finally in Section V we conclude the paper.

## II. Related Works

In this section the most relevant works to this study are reviewed. The focus of the review has been put on the works related to the computation linguistic, machine learning and statistical-based classifier. The review is conducted in a systematic way. That is, five prominent aspects of the reviewed works have been considered. The aspects are: 1. aim of the work, 2. scope of the work, 3. features extracted, 4. classifiers type, and 5. success ratio.

A study conducted by (Cheng, et al., 2009) for identifying the gender of email composers. The scope of the work was short length, multi-genre, content-free emails. The authors extracted 545 features distributed over five classes. The Features were character, word, syntactic, paragraph and function based. The authors used decision tree and support vector machine classifiers to identify email owners, and they obtained an average accuracy of 82.2%. The same previous research team conducted another set of experiments (Cheng, Chandramouli and Subbalakshmi, 2011). This time they used Reuter's newsgroup dataset, a collection of reports and articles form journalists in English language. The authors augmented the Reuter's dataset with the informal email texts in order to have a mixed dataset of formal and informal English writing. During the study, they extracted 545 features of five types. They also used three types of classifiers, decision tree, support vector, and machine and Bayesian-based logistic regression classifier. They obtained this time an accuracy rate of 85%. The study concluded that among the five feature types studied, the word, structure and function features were more effective than the other two features, paragraph and syntactic.

Texts collected from the Tweeter and other social networks and blogs comprise a good experimental dataset. This is because there is no censorship on these social networking and people who use this service are entirely free to express what they like/dislike. However, there is a problem associated with using social networking data. That is, labeling and parsing account owners need to be performed manually before they can be used for training set.

A research has been conducted by (Burger, et al., 2011) to discriminate gender of account owners in the social network Tweeter. Features extracted in Burger's study were n-gram for both word-level and character-level. The extraction process done for four types of text, the profile name, full name, description and tweets. Each extracted n-gram feature is a simple indicator of zero and one, one for existing and zero is the future doesn't exist. The authors used support vector machine, Naive Bayes and Balanced Winnow2 classifiers, and they obtained an accuracy ratio 67% for Naïve Bayes, 71% for SVM and 74% for Winnow2. A noticeable difference between this study and other previous studies is this one worked on a collection of languages instead of only English language used by others.

Another group of researchers (Deitrick, et al., 2012) have conducted a study to identify the gender of account owners in Tweeter. What is unique about this study is that the authors accounted for two important aspects of the social networking. That is, they considered steam processing, and feature reduction. The first aspect is important because text/data changes and get updated with the time and the second aspect is conducted as a preprocessing step for performance improvement.

They extracted about 9000 features of type n-gram (1-gram and 2-gram) where n-gram refers to an existence/inexistence of a character or two in a passage. In Addition to using 1-gram and 2-gram features, the authors also used the feature selection process, also called also called dimensionality reduction, which is about selecting the most affective features among all features to improve efficiency in large data sets. In the study a special kind of neural networks classifier called Modified Balanced Winnow classifier is used for processing streams. The study obtained 82.48% accuracy. This rate increased to 98.5% after feature reduction process.

A research done by (Herdağdelen, 2013) was concentrated on a potential combining of an n-gram text corpus from twitter messages with demographics metadata. He used these messages coupled with metadata about their authors to understand a wide variety of phenomena ranging from political polarization to geographic and demographic lexical variation. Gender was among these metadata used with n-gram for this purpose.

Another team (Nguyen, et al., 2011) has studied text for other human properties but not the gender. They investigated manifest properties of textual messages, including latent topics, psycholinguistic features, and author mood, of a large corpus of blog posts, to analyze the impact of age, emotion, and social connectivity. These properties are found to be significantly different across the examined cohorts. They build binary classifiers for old versus young bloggers, social versus solo bloggers, and happy versus sad posts with high performance.

## III. Challenges and Assumption

This section briefly describes gender identification challenges for online newspapers and our assumptions as well as the difficulty we faced while processing text writing in Kurdish language.

### *A. Challenges in Gender Identification*

Identifying true genders of authors from writing style faces two inherited challenges: the first one is Editing, and the second one is Unified writing style.

Site owners, similar to paper based newspapers, may edit text or apply templates to unify writing style on their sites, thus, intervening in the authors' writing style. For the purpose of this study, we assume that site owners publish articles without editing. This assumption is valid for Kurdish online

newspapers, because most of these newspapers publish their service free of charge, and they do not employ professional editors. To address the second problem, we choose long and medium size articles for our study. It has been proven that despite applying templates in order to unify writing styles the writing style in long and medium size articles remain unchanged.

Using long passages have helped us to avoid, yet another problem which we named *disguise problem*, female authors intentionally following male styles of writing. In Kurdistan, journalism is a male dominate professional where majority of the famous journalists, reporters, analyzers are males. Hence, sometimes Kurdish female authors intentionally follow male styles writing, thus, making author gender identification to be even a harder problem. This situation is minimized using long passages in our study. This is because it is harder for female authors to continue act like males in long writings.

### *B. Kurdish Text Processing Challenge*

Kurdish language has two writing script, Latin and Arabic based writing scripts. In this study the focus is put on the Arabic based writing script. Using Arabic based script for our study is not trouble free. It requires special IDE configuration and programming language knowledge in order to process text. This problem is solved by using techniques and programs used by (Yunis, 2012).

## IV. The Classifier Model and Experimental Results

Extracting effective features from text is the most important step for discriminating between genders in the authors writing style. This section explains preliminary steps and experimental setups, then lists the features extracted for this study followed by a description for the classifier model used in this study and finally results are presented. Generally the system can be illustrated in the process diagram shown in the Fig. 1.

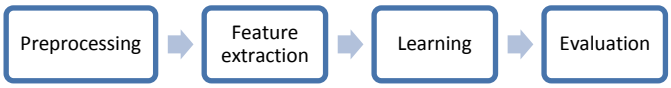


Fig. 1 Process diagram of the system

### *A. Preliminary Steps and Experimental Results*

For the purpose of this study four online Kurdish newspapers were selected. These newspapers are Hawlati, Khabat, Hawler, and Gulan. The basis for selecting these newspapers was their popularity; the paper-based versions of these newspapers are distributed across Kurdish region. The total number of articles selected for the experiment was 75. Each article is labeled by male or female (according to the name of the journalist). Table I presents the number of articles taken from each newspaper as well as number of male/ female authors in each newspaper.

In order to keep the originality and avoid losing writing style, no editing or cleaning processes were executed on the collected data. After tokenizing the collected data according to spaces, the number of words in each article is counted. The result of this step is depicted in Table II.

TABLE I
Number of Articles According to Newspapers and Gender of Journalists

| Newspaper | Number of articles | Number of males | Number of females |
|---|---|---|---|
| Hawlati | 6 | 5 | 1 |
| Khabat | 29 | 20 | 9 |
| Hawler | 16 | 13 | 3 |
| Gulan | 24 | 17 | 7 |

TABLE II
Average Number of Words According to Newspapers and Gender of Journalists

| Newspaper | Average number of words | Average number of words (male) | Average number of words (female) |
|---|---|---|---|
| Hawlati | 802 | 828 | 696 |
| Khabat | 896 | 841 | 1018 |
| Hawler | 374 | 417 | 184 |
| Gulan | 859 | 963 | 606 |

### *B. Feature Extraction*

Each passage has its own embedded features, and existing gender identification solutions are based on feature extraction. Hence, this step is considered as the most effective step out study. For this study, 61 unique features were extracted for each article. The extracted features were clustered into five different types. These feature types are character, word, syntactic, paragraph and 1-gram based features. All extracted features are showed in the Table III.

A Java code was used to extract features from articles because Java has the ability to process Kurdish texts easily (Yunis, 2012). Fig. 2 is the algorithm used for extracting number of letters in an article as one of the character-based features. Fig. 3 is the algorithm used for extracting number of words in an article as one of the word-based features. Fig. 4 is the algorithm used for extracting number of question marks in an article as one of the syntactic-based features. Fig. 5 is the algorithm used for extracting number of sentences in an article as one of the paragraph-based features. Fig. 6 is the algorithm used for determining the frequency of letter " د " in an article as one of the 1-gram features.

The extracted features are collected in a Microsoft- Excel sheet to be processed by a neural model. Fig. 7 shows a sample of the features extracted from articles to be used as a training set later.

For classifying purposes and training a supervised model target classes are necessary. The target classes are collected in a Microsoft-Excel file sheet, this sheet is processed later by the model. Fig. 8 shows a sample of the target classes, as mentioned before the classes are labeled manually.

The same spread sheets (features sheet and targets sheet) are prepared for the evaluation set, in other words, two other Microsoft-Excel sheets are prepared for determining the accuracy rates. This step is better explained in section IV-G.

TABLE III
EXTRACTED FEATURES

| Feature | Number of extracted features | Features |
|---|---|---|
| Character-based | 8 | Total number of special characters<br>Total number of letters<br>Total number of special characters and letters<br>Total number of special characters, letters and spaces<br>Total number of white space<br>Ratio of letters over special characters<br>Ratio of letters over characters<br>Ratio of letters over special characters and letters |
| Word-based | 6 | Total number of words in the article<br>Ratio of word length<br>Words longer than six characters<br>Average words longer than six characters<br>Total number of short words<br>Average number of short words |
| Syntactic-based | 12 | Number of commas<br>Ratio of commas to characters<br>Number of periods<br>Ratio of periods to characters<br>Number of colons<br>Ratio of colons to characters<br>Number of semicolons<br>Ratio of semicolons to characters<br>Number of question marks<br>Ratio of question marks to characters<br>Number of exclamation marks<br>Ratio of exclamation marks to characters |
| Paragraph-based | 2 | Number of paragraphs.<br>Number of sentences |
| 1-gram | 33 | Frequency of letter "ا"<br>Frequency of letter "ب" , and all other Kurdish letters. |

### C. The Software Used for Training

The software used in this paper for learning the model is MATLAB 7.6. It provides a very easy to use and friendly environment which makes our job easier. An important note here is that MATLAB initializes all weights and bias values required to start learning process. Users are only asked to specify the structure of the model like number of hidden layers and number of nodes inside each layer, also type of the transfer functions for each layer should be specified. The extracted features and the target classes, which are already saved in the Excel sheets, are fed to the model as parameters.

```
algorithm: Total number of letters in an article
input: string article
output: integer i

for all line ∈ article do
    for all word ∈ line do
        if word is not punctuation then
            for all letter ∈ word do
                if letter is not punctuation then
                    increment i
return i
```

Fig. 2. The algorithm for extracting number of letters in an article.

```
algorithm: Total number of words in an article
input: string article
output: integer i

for all line ∈ article do
    for all word ∈ line do
        if word is not punctuation then
            increment i
return i
```

Fig. 3. The algorithm for extracting number of words in an article.

```
algorithm: Total number of question marks in an article
input: string article
output: integer i

for all line ∈ article do
    for all word ∈ line do
        for all letter ∈ word do
            if letter == " ? " then
                increment i
return i
```

Fig. 4. The algorithm for extracting number of question marks in an article.

```
algorithm: Total number of sentences in an article
input: string article
output: integer i

for all line ∈ article do
    for all word ∈ line do
        for all letter ∈ word do
            if letter == " . " then
                increment i
return i
```

Fig. 5. The algorithm for extracting number of sentences in an article.

```
algorithm: Total number of letter " د " in an article
input: string article
output: integer i

for all line ∈ article do
        for all letter ∈ line do
                if letter == " د " then
                        increment i
return i
```

Fig. 6. The algorithm for determining the frequency of letter " د ".

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 11 | 2884 | 2895 | 3315 | 420 | 262 | 0.869984917 | 0.996200345 | 416 | 6.932692308 | 218 |
| 2 | 8 | 2220 | 2228 | 2632 | 404 | 277 | 0.843465046 | 0.996409336 | 402 | 5.52238806 | 139 |
| 3 | 12 | 4472 | 4484 | 5227 | 743 | 372 | 0.855557681 | 0.997323818 | 743 | 6.01884253 | 314 |
| 4 | 19 | 4664 | 4683 | 5464 | 781 | 245 | 0.853587116 | 0.995942772 | 782 | 5.964194373 | 310 |
| 5 | 3 | 3401 | 3404 | 4095 | 691 | 1133 | 0.830525031 | 0.999118684 | 692 | 4.914739884 | 205 |
| 6 | 9 | 3527 | 3536 | 4151 | 615 | 391 | 0.849674777 | 0.997454751 | 615 | 5.73495935 | 246 |
| 7 | 6 | 9318 | 9324 | 10857 | 1533 | 1553 | 0.858248135 | 0.999356499 | 1534 | 6.074315515 | 625 |
| 8 | 17 | 4897 | 4914 | 5762 | 848 | 288 | 0.849878514 | 0.996540497 | 849 | 5.767962309 | 318 |
| 9 | 35 | 6816 | 6851 | 7818 | 967 | 194 | 0.871834229 | 0.994891257 | 967 | 7.04860393 | 505 |
| 10 | 14 | 4910 | 4924 | 5594 | 670 | 350 | 0.877726135 | 0.997156783 | 669 | 7.339312407 | 364 |
| 11 | 26 | 2562 | 2588 | 3014 | 426 | 98 | 0.850033179 | 0.989953632 | 427 | 6 | 163 |
| 12 | 14 | 4311 | 4325 | 5162 | 837 | 307 | 0.835141418 | 0.996763006 | 838 | 5.144391408 | 226 |
| 13 | 19 | 9106 | 9125 | 10836 | 1711 | 479 | 0.840346992 | 0.997917808 | 1712 | 5.318925234 | 491 |
| 14 | 13 | 7363 | 7376 | 8771 | 1395 | 566 | 0.839470984 | 0.998237527 | 1395 | 5.278136201 | 426 |
| 15 | 35 | 5655 | 5690 | 6684 | 994 | 161 | 0.846050269 | 0.993848858 | 995 | 5.683417085 | 348 |
| 16 | 24 | 4539 | 4563 | 5350 | 787 | 189 | 0.848411215 | 0.994740302 | 778 | 5.834190231 | 315 |
| 17 | 32 | 11962 | 11994 | 14013 | 2019 | 373 | 0.85363591 | 0.997331999 | 2019 | 5.924715206 | 811 |

Fig. 7. The features collected in an MS-Excel sheet.

| | A | B | C | D |
|---|---|---|---|---|
| 1 | 1 | | | |
| 2 | 1 | | | |
| 3 | 1 | | | |
| 4 | 1 | | | |
| 5 | 1 | | | |
| 6 | 1 | | | |
| 7 | -1 | | | |
| 8 | -1 | | | |
| 9 | -1 | | | |
| 10 | -1 | | | |
| 11 | -1 | | | |
| 12 | -1 | | | |
| 13 | -1 | | | |
| 14 | -1 | | | |
| 15 | -1 | | | |
| 16 | -1 | | | |

Fig. 8. The target classes collected in an MS-Excel sheet.

### D. Data Normalization

Data normalization is not necessary in MLP because input variables are combined linearly, then it is rarely strictly necessary to standardize the inputs. For making sure the input set is normalized and given to the model, there was no sensible difference between results. The reason is that any rescaling of an input vector can be effectively undone by changing the corresponding weights and biases, getting the outputs as you had before. There are a variety of practical reasons why standardizing the inputs can make training faster and reduce the chances of getting stuck in local optima but because that the training algorithm is Levenberg-Marquardt, and it's known that LM is very fast and it can overcome all local minimums, for these reasons the data were not normalized but fed directly to the model.

### E. The Classifier

The technique used in this paper is a Multi-layer Perceptron (MLP) Neural Network Binary Classifier model. This is because MLP model is a very efficient classifier for binary classification (in our case male vs. female). This MLP model consists of three layers: an input, a hidden layer and an output layer. The number of nodes in the input is 61, each node is specialized for inputting an extracted feature. The number of nodes in the hidden layer is 30, and there is only one node in the output layer. The structure can be summarized as 61-30-1. The model structure is illustrated in the Fig. 9.

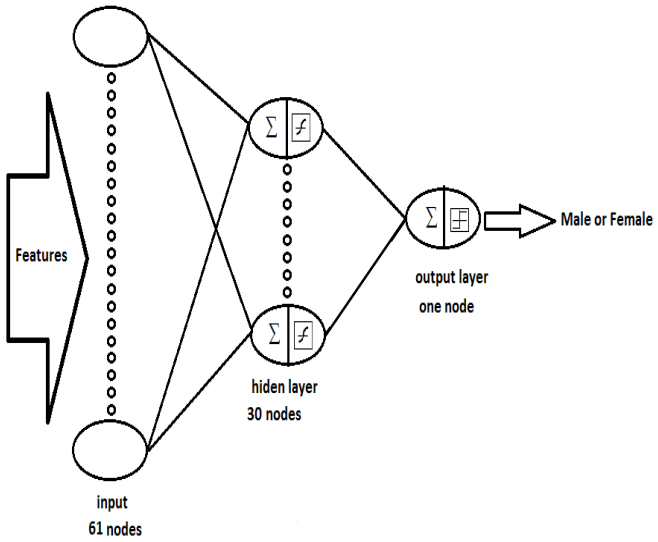


Fig. 9. Multi-layer perceptron neural network model.

The information in MLP moves forward (feed-forward) which is a simplest type of neural networks, in this type of network the information moves from input nodes to the hidden layer nodes and finally to the output layer nodes, information is not going back in a cyclic path to the input nodes or hidden layer nodes. A single node in the neural network model is called perceptron, a model consist of a number of these perceptrons arranged in layers that is why sometimes called multi-layer perceptron or MLP. A multi-layer perceptron or feed-forward network may consist of a single-layer or more than one layer. In such networks number of hidden layers is optional. Notice that the word "layer" hasn't been appended to the word "input". This is because input is not a real layer (no summation, no bias, and no transfer function). For explaining how the network used in this paper works, it's necessary to understand how a simple perceptron or a neuron works. Let us take a neuron from hidden layer as an example, Fig. 10.

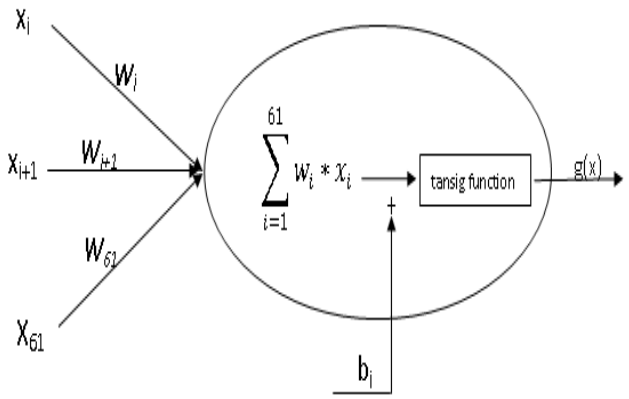


Fig. 10. A simple neuron from hidden layer

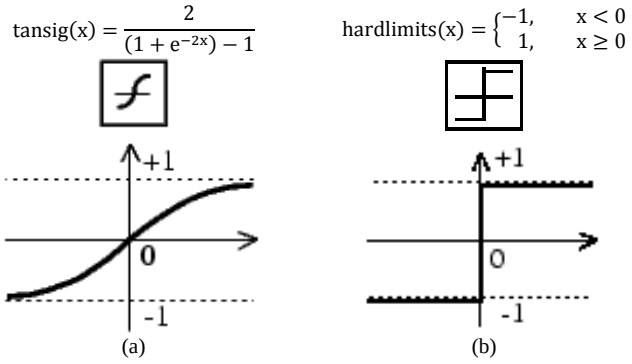


Fig. 11. (a) Hyperbolic tangent function. (b) Symmetric hard limit function

The information that fed to the neuron from 61 input neurons are multiplied by their connection weights then summed together, also summed with the bias value. The resultant value is fed to the transfer function in this case a tansig, Fig. 11-a, the result of this function is the output of the neuron or perceptron. The resultant value will be between 1 and -1, because of the range of tansig function which lays in this period. (1) gives the feed-forward mathematical process in the hidden layer:

$$d(j) = tansig\left\{ \quad \left[\sum_{i=1}^{61} \mathrm{X(i)} * \mathrm{W(i)}\right] + [b(\mathrm{j})] \quad \right\} \qquad (1)$$

where: $d$(j) is the output of a neuron in the hidden layer, $X$(i) is the input value from input neuron i, $W$(i) is the weight of connection between input neuron i and the neuron j, and $b$(j) is the bias value of the neuron j.

The same process is repeated in all 30 neurons of the hidden layer, then this 30 information are fed to the output layer and the same process is repeated. Number of nodes in the output layer is only a one node. The transfer function here is different, symmetric hard limit function was used as in Fig. 11-b, where the result takes only two values -1 or 1. Class 1 is used for female whereas class -1 is for male. The feed-forward mathematical process in the output layer is declared in (2);

$$D = symmetrichardlimits\left\{\left[\sum_{\mathrm{j}=1}^{30} X(\mathrm{j}) * W(\mathrm{j})\right] + [b]\right\} \qquad (2)$$

where: $D$ is the output of a neuron in the output layer, $X$(j) is the output of the neuron j from hidden layer, $W$(j) is the weight of connection between hidden neuron j and the output neuron, and $b$ is the bias value of the output neuron.

### *F. Learning Process*

Back-propagation is a common method used for supervised learning. In this method an artificial network learns from comparing an output with desired outputs, then propagating the error occur in a backward direction by justifying the weights of the connections between nodes. This backward propagation needs the transfer functions used in the nodes to be differentiable to ensure smooth back-distribution of errors on the weights.

In the supervised learning, desired or target results is compared with the obtained results and the squared error is calculated according to (3):

$$Er = (T - D)^2 \qquad (3)$$

where: $Er$ is a squared error, $T$ is the target value, and $D$ is the desired value

Optimization methods are used to minimize this error value $Er$. There are lots of methods that can be used for this issue. In this paper a Levenberg-Marquardt optimization is used for optimizing errors. The LM is one of the best and most efficient methods for optimizing backpropaation of errors which designed to approach second-order training speed (Esugasini, et al., 2005). Not like other backpropagation algorithms, LM is very fast and doesn't have a problem of local minimums.

### *G. Results*

During the study, 65 of the selected articles (87%) were used for training the neural network model, and the other 10 articles (13%) were used for evaluation process. The model tested 10 times over different sets, randomly separated to training and evaluation sets (i.e each time 65 for training and 10 for evaluation). The average of the 10 trials was taken in consideration as accuracy of the model. Table IV summarizes the experimental results obtained at each test as well as the average accuracy rate for the study, 76%.

TABLE IV
THE ACCURACY RATIO OF 10 FOLDS

| **Fold** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Accuracy %** | 60 | 80 | 90 | 70 | 80 | 80 | 80 | 70 | 70 | 80 |
| **Average %** | 76 | | | | | | | | | |

The result in Table IV indicates that if our model is used for determining or discriminating the gender of a journalist or a column writer in a web-based Kurdish media the model can obtain result with the 76% accuracy rate.

## V. Conclusion

Discriminating between genders through writing styles is a difficult problem in Kurdish language. This is because the language itself (Sorani dialect) doesn't discriminate between genders. This work is a first step towards a comprehensive study on the gender identification using writing styles in Kurdish language. The study used online newspapers as a data source, and feed-forward MLP as the classifier. During the study, 61 unique features of five types were extracted. These features types were character, word, syntactic, paragraph and 1-gram based features. The accuracy rate of the study was 76% success rate. Tested samples contained persons have good knowledge in punctuation marks the system will identified them too. Also it has been concluded that the accuracy rate various from one language to another, and one of the reasons for the difference among languages is due to the language flexibility to express gender differences and emotions.

## References

Burger, J., Henderson, J., Kim, G. and Zarrella, G., 2011. Discriminating gender on Twitter. In: Association for Computational Linguistics, C*onference on empirical methods in natural language processing*, 27-31 July 2011. Edinburgh, Scotland, UK.

Cheng, N., Chandramouli, R. and Subbalakshmi, K.P., 2011. Author gender identification from text. *Digital Investigation*, 8(1), pp.78-88.

Cheng, N., Chen, X., Chandramouli, R., and Subbalakshmi, K., 2009. Gender identification from e-mails. In: IEEE, *IEEE Symposium on computational linguistics and data mining proceedings*, 30-2 April 2009. Nashville, TN, USA.

Deitrick, W., Miller, Z., Valyou, B., Dickinson, B., Munson, T. and Hu, W., 2012. Gender identification on Twitter using the modified balanced winnow. *Communications and Network*, 4(3), pp.189-195.

Efron, R., and Thisted, B., 1976. Estimating the number of unseen species: How many words did Shakespeare know?. *Biometrika*, 63(3), pp.435-447.

Esugasini, S., Mashor, M., Isa, N. and Othman, N., 2005. Performance comparison for MLP networks using various backpropagation algorithms for breast cancer diagnosis. In: *9th International conference on knowledge-based intelligent information and engineering systems* (KES'05), 14-16 September 2005. Australia.

Herdağdelen, A., 2013. Twitter n-gram corpus with demographic metadata, *Language resources and evaluation*, pp.1-21.

Labov, W., 1990. The intersection of sex and social class in the course of linguistic change. *Language variation and change*, 2, pp.205-254.

Lakoff, R., 1973. Language and women's place. *Language in society*, 2(1), pp.45-80.

Merriam, T., 1996. Marlowe's hand in Edward III revisited. *Literary and linguistic computing*, 11(1), pp.19-22.

Nguyen, T., Phung, D., Adams, B. and Venkatesh, S., 2011. Prediction of age, sentiment, and connectivity from social media text. In: WISE (Web Information System Engineering), *12th International conference on web information system engineering* (WISE'11), 12-14 October 2011. Sydney, Australia.

Yunis, A. M., 2012. *Towards an application programming interface (API) for processing Kurdish text*. [pdf] Canada: Carlton University research group web-site, Available at: http://people.scs.carleton.ca/~armyunis/projects/KAPI/KAPI.pdf