Predicting the Gender of the Kurdish Writers in Facebook

Peshawa J. Muhammad Ali

Department of Engineering Software
Koya University

Received : 5/7/2013
Accepted : 17/11/2013

Abstract

Facebook is one of the social networks which have lots of users among Kurdish people. Although there are no formal or published statistics about the number of the Facebook users, in the last few years Facebook was the most used website among Kurdish society. This swift development of the Kurdish society towards Facebook imposes new challenges that need to be addressed. For example, a poem or an article published on Facebook possesses properties such as author name, gender, age, and nationality among others. In this paper the gender of Kurdish authors in Facebook determined by using a feed-forward artificial neural network model. 120 Facebook Kurdish written posts were used for learning the model designed to determine the gender of Kurdish writers in Facebook. The posts were taken from Facebook pages of different persons with different backgrounds. Twenty eight text features were extracted from each post; these features were distinct in discriminating between genders. The feed-forward back-propagation artificial neural network with three layers (28 nodes, 14 nodes, 1 node) is used as a classification technique. The accuracy ratio which based on the ten-fold technique (taking the average ratio among ten trials) obtained was 77.5 %. This proposed idea of this paper is important for detecting the real gender of Facebook page owners.

Keywords: Facebook, neural networks, text mining, gender identification.

1. Introduction

Facebook community is one of the biggest communities in the world that grows rapidly. According to the Facebook's newsroom (https://newsroom.fb.com/Key-Facts) there are 1.15 billion monthly active users as of June 2013. It's more attractive to the Kurdish users than all other social networks. This attraction has its impact on all other parts of life in this society; it makes people express their feelings, emotions, etc easily and without restriction. The role of Facebook was important in the revolutions happened in the Middle East countries generally. According to Kurdistan Region's Ministry of Communication and Transportation (www.moc-krg.com) there are two major companies that provide internet services and more than 20 smaller companies providing internet services. However, this growth encourages journalists, poets, reporters, analyzers, politicians,...etc. to publish their thoughts, opinions, reports and analysis more easily and even without any restrictions. Among these web-based communities there are large numbers of women participating actively in these changes. Hence, gender identification becomes a significant issue.

Professional presses follow a kind of publishing regulations and ethics while this restriction is not exist in Facebook. Any person can write a post or a partial idea on his/her own page. This is because an account can be created easily and can share what he\she likes without proper review. This encourages a kind of misuse by broadcasting deceptive and false information. Fake personalities on social networks like Facebook and Tweeter are widespread phenomena. So, the problem is that a person can create a page and write what he/she likes without any constraints and hide his/her personality. One of the things that we would like to know is that “is the author is male or female?". Moreover, these kinds of information can be extracted from the post itself. More precisely, the main question of this work is that “Is it possible to distinguish male writers from female writers based on their writings?”, if the answer is yes for languages such as English and Arabic, is this applicable to Kurdish language? This paper will examine this issue. Identification of gender will help partially in solving the problem of faked personalities.

Authorship identification is a wider research area than gender identification. Studies in this area include the attribution of disputed Shakespearian poems done by Efron and Thisted [1], and Merriam [2]. Nonetheless, gender identifying researchers examined a specific part of authorship identification such as Lakoff [3], and Labov [4]. They all stated that there is a difference in the style of writing for male and female writers. In all these works the most important thing they concentrated on is feature extraction. Each passage written by an author has its features embedded in the text itself. Our job is to extract these features to recognize the gender of the author.

Throughout published papers, the scope of the research done by Cheng N et al [5] was email texts only. The classifiers were Support Vector Machine, and Decision Tree. The scope of the research done by Cheng N. et al [6] was a collection of journal reports and informal emails. The classifiers were Decision Tree, Support Vector Machine and Bayesian-based Logistic Regression. Both researches classified the features to character-based features, word-based features, syntactic features, structural features, and function words in which a human may use some words more than others.

The classification used by a research done by Burger et al [7] were Support Vector Machine, Naive Bayes and Balanced Winnow2, while the classification used a research done by Deitrick et al [8] were Modified Balanced Winnow and a special kind of neural networks used for processing streams. The scope of both works was the tweets from the social network Tweeter. Both researches concentrated on extracting n-gram features, which was a statistical based feature extraction used for discriminating between authors' gender.

Informal articles like blogs and social network posts may be more recognizable than formal writing such as web-based media, why? Because, in web-based formal media there is a possibility of site editor interventions, that is, sometimes site editors may fix some linguistic and spelling mistakes or they put unified writing styles. This possibility in Facebook is avoided because there is no intervention, the owner can freely write. Another problem is that sometimes authors especially females, follow male styles in writing, or repeat male styles; especially in Kurdish society because most of the famous poets, journalists, reporters, analyzers are males. This problem can be avoided by knowing that even if a female tried to act like a male, she can't continue throughout the passage, so the length of the article will be enough to uncover them. Therefore, authors can be detected more precisely in social networks and blogs.

2. The proposed approach

The proposed approach can simply divide into five steps, Figure (1) illustrates the block diagram of the whole system.

2.1. Data gathering

The 120 Facebook posts from different accounts were collected. They all belong to Kurdish writers. Some of them are very famous writers, others are just beginners. The collection contains parliamentarians, poets, actors, politicians, singers, and young and old people from both genders. Therefore, the scope of this study is the Kurdish writers on Facebook. Each article is labeled by male or female manually (according to the name of the authors). Table (1) explains the number of articles taken from Facebook as well as the number of males and females. No editing or cleaning processes are executed on the collected data because articles may loss its characteristics if any kind of parsing processes applied. After tokenizing the collected data according to spaces, the number of words in each post is counted and showed in the Table (2).

2.2.Features extraction

This step is very important because the types of the extracted features will affect the process of discrimination. Twenty eight text features are extracted from each post, divided into four types of text features. These features are paragraph-based features illustrated in the Figure (2), character-based features illustrated in the Figure (3), word-based features illustrated in the Figure (4), and syntactic-based features illustrated in the Figure (5). All extracted features are shown in the Table (3).

2.3. Model creating

A feed-forward neural network is a very efficient classifier especially for binary classification (i.e. 0

vs. 1, white vs. black and male vs. female). It consists of a number of neurons arranged in layers. Each neuron is called perceptron. The feed-forward neural network is a network of perceptrons with the connecting links characterized by weights. Mathematically, a neuron is given in equation (1), and Figure (6) illustrates a simple perceptron neuron.

$$y = Ft\{[\sum_{i=1}^{n}(x_i * w_i)] + b_j\} \qquad (1)$$

where x1,.....,xn are input features, w1,.....,wn are the weights of the connections, b_j is the bias value, and Ft is a transfer function, and y is the output of this single neuron.

In this paper, a feed-forward neural network model is used or sometimes called multi-layer perceptron (MLP) which consists of an input, a hidden-layer and an output-layer. Notice that there is no layer word appended to the word input because the input is not a real layer (no summation, no bias, and no transfer function). The number of nodes in the input is 28 nodes equal to the number of features while the number of nodes in the hidden layer is 14 nodes and only one node in the output layer. The structure is 28-14-1 as illustrated in the Figure (7).

The transfer function used for hidden layer is called tansig function which is the hyperbolic tangent sigmoid function. The transfer function used for the output layer is a hard limit function or a hardlim. The mathematical expressions of the transfer functions are explained in equations (2) and (3).

$$tansig(x) = \frac{2}{(1+e^{-2x})-1} \qquad (2)$$

$$hardlim(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \qquad (3)$$

Figure (8) explains the sketches of the transfer functions and signs used for both of them. The tansig transfer function is mathematically equivalent to tanh, but it runs faster in MATLAB than tanh with very small numerical differences. This function is a good tradeoff to be used instead tanh in neural networks, where speed is important and the exact shape of the transfer function is not.

2.4. Model training

For training a feed-forward neural network a backpropagation algorithm is used which is an abbreviation for "backward propagation of errors". It is a common method for training artificial neural networks. The model learns from desired outputs. It compares the obtained results with the desired ones then it tries to propagate the difference between them in a backward way on the weights. Weights and biases are updated according to this new situation then another cycle of a feed-forward is computed then another backpropagation till the computed error between real output and desired output reaches a limit that can be accepted.

It is a supervised learning method, it requires a dataset of the desired output for many inputs, making up the training set. Backpropagation requires the transfer function used by the artificial neurons which is differentiable. In this paper, a Levenberg-marquardt (LM) algorithm is used. It provides a numerical solution to the problem of minimizing a function, generally nonlinear, over a space of parameters of the function.

120 posts were labeled by male and female, 108 of these articles (90 %) were used for training the neural network model while the other 12 (10 %) posts were used for evaluating the process. MATLAB was used for creating and training the model.

3.Experimental Results

A ten-fold method is used for determining the accuracy of the model. The ten-fold is a cross-validation process used for evaluating classification performance for neural network models and other decision making techniques. In this process, models tested 10 times over different sets, randomly separated into training and evaluation sets (i.e. each time 108 for training and 12 for evaluation) and then the accuracy ratio is calculated. The accuracy is the ratio of the correct gender determination samples to all tested samples. According to the procedure of ten-fold the average of all ten folds are taken into consideration. Resultant ratios of all folds are explained in the Table (4).

4.Conclusion

Texts hold features related to their authors like age or gender of the author. Gender of the Facebook account owners can be discriminated from their styles of writing. Artificial neural network were used successfully for classifying the gender of Kurdish account holders from their

posts. The performance of the model was 77.5%. The model is working better with casual writings like social networks more than formal texts like books or journals, because they are subjected to a kind of editing regulations from editors which make the articles loss features of their author this case is not exists in Facebook. There are languages giving more freedom to authors to express their emotions and to have their styles of writing, Kurdish hasn't a high flexibility for this purpose.

References

1. Efron R, Thisted B. "estimating the number of unseen species: How many words did Shakespeare know?", Biometrica, 1976, 63(3), 435-447.
2. Merriam T. "Marlowe's hand in Edward III revisited", Literary and linguistic computing, 1996, 11(1), 19-22.
3. Lakoff R.T, "Language and women's place", Harper Colophon Books, New York, 1975.
4. Labov W. "The intersection of sex and social class in the course of linguistic change", Language variation and change, 2, 1990.
5. Cheng N, Cheng X, Chandramouli R, Sabbalakshmi K.P, "Gender identification from e-mials", IEEE symposium on computational linguistics and data mining proceedings, 154-158, 2009.
6. Cheng N, Chandramouli R, Sabbalakshmi K.P, "Author gender identification from text", Digital investigation, 8, 78-88, 2011.
7. Burger J. D, Henderson J, Kim G, Zarella G, "Discriminating gender on Tweeter", Technical report, Mitre Corporation, Bedford, Massachusetts, USA, 2011.
8. Deitrick W, Miller Z, Valyou B, Dicknison B, Munson T, Hu W; "Gender Identification on Twitter using the modified balanced winnow", Communications and network, 4, 189-195, 2012.

خەملاندنی ڕەگەزی نوسەری کورد لە فەیسبووك

پێشەوا جمال محمد علی / ماموستا یاریدەدەر
بەشی ئەندازیاری سوفت ویر ، زانکۆی کۆیە

فەیسبووك یەکێکە لە تۆڕە کۆمەلأیەتییە بەناوبانگەکان ، لە ناو خەڵکی کوردستانیشدا بەکارهێنەرێکی زۆری هەیە. لەو چەند ساڵەی دوایی دا تێبینی دەکرێت کە فەیسبووك ئەو وێبسایتە بووە کە زۆرترین کەس لە کوردستاندا سەردانیان کردووە. ئەو گەشەسەندنە خێرایە هەندێك ڕەهەندی نوێی بەدوای خۆیدا هێنا بۆ نموونە کاتێك پۆستێك یان نوسینێك لەسەر لاپەڕەی فەیسبووکی تایبەتی کەسێك بلأودەکرێتەوە، هەڵگری هەندێك سیفاتی خاوەنی نووسینەکەیە کە دەتوانرێت بدۆزرێتەوە وەك جێندەر یان تەمەنی کەسی نووسەر یان ئایا خەڵکی چی دەڤەرێکە.لەو توێژینەوەیەدا هەستاوین بە دیاری کردنی ڕەگەزی ئەو کەسەی کە پۆستێك بلأو دەکاتەوە لە فەیسبووك لە ڕێگای شێوازی نووسینەکەیەوە بە بەکارهێنانی تەکنیکە ژیریەکان وە هەڵێنجانی زانیاری لە تێکست. هەستاین بە کۆکردنەوەی 120 پۆستی جیاواز کە هی کەسانی جیاواز بوون لە شیعر و نووسینی کوردی، 28 خاسیەتی جیاواز لە هەریەکە لەو بابەتانە دەرهێنران، وە مۆدێلکی ژیری تایبەت کە(نیورەڵ نیتۆرك- Neural network) تیایدا بەکارهێنرابوو ئامادەکرا. دواتر مۆدێلەکە فێرکرا کە هەستێت بە جیاکاری لەنێوان ڕەگەزی نوسەر لەسەر بنەمای نووسینەکەی واتا ئایا نووسەر نێرە یان مێ یە؟ ڕێژەی دروستی خەملاندنەکان بریتی بوو لە 77.5 % (ڕێژەی ناوەند لە دە جار دا بە ڕێگای Ten-fold). ئەو توێژینەوەیە گرنگە بۆ ئاشکرا کردنی ئەو کەسانەی کە لە تۆڕە کۆمەلأیەتی یەکاندا هەڵدەستن بەگۆڕینی کەسایەتی خۆیان وە لەوانەیە چەندین کاری ساختەکاری پێ ئەنجام بدەن.

◆

Table (1): Number of articles and poems according to the gender of the writers

Posts	Gender	
	Males	Females
Poems	30	20
Articles	40	30

Table (2): Average number of words according to the type of the post and gender of journalists

Posts	Average number of words	Average number of words (male)	Average number of words (female)
Poems	72	65	81
Articles	288	304	257

Table (3): Extracted features

Feature	Number of extracted features	Features
Paragraph-based	2	-Number of paragraphs. -Number of sentences.
Character-based	8	-Total number of special characters. -Total number of letters. -Total number of special characters and letters. -Total number of special characters, letters and spaces. -Total number of white spaces. -Ratio of letters over special characters. -Ratio of letters over characters. -Ratio of letters over special characters and letters.
Word-based	6	-Number of words less than six characters. -Number of words more than or equal to six characters. -Total number of words in the article. -Average number of characters per word. -Average number of characters per word for long words. -Average number of characters per word for short words.
Syntactic-based	12	-Number of commas. -Ratio of commas to characters. -Number of periods. -Ratio of periods to characters. -Number of colons. -Ratio of colons to characters. -Number of semicolons. -Ratio of semicolons to characters. -Number of question marks. -Ratio of question marks to characters. -Number of exclamation marks. -Ratio of exclamation marks to characters.

Table (4): The accuracy ratio of 10-folds

Fold	1	2	3	4	5	6	7	8	9	10
Accuracy %	66.66	75	83.33	75	83.33	66.66	75	91.66	83.33	75
Average %	77.5									

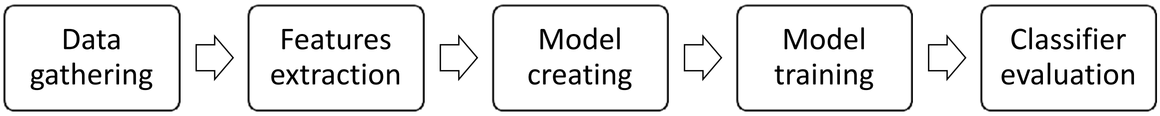

Figure (1): The block diagram of the system

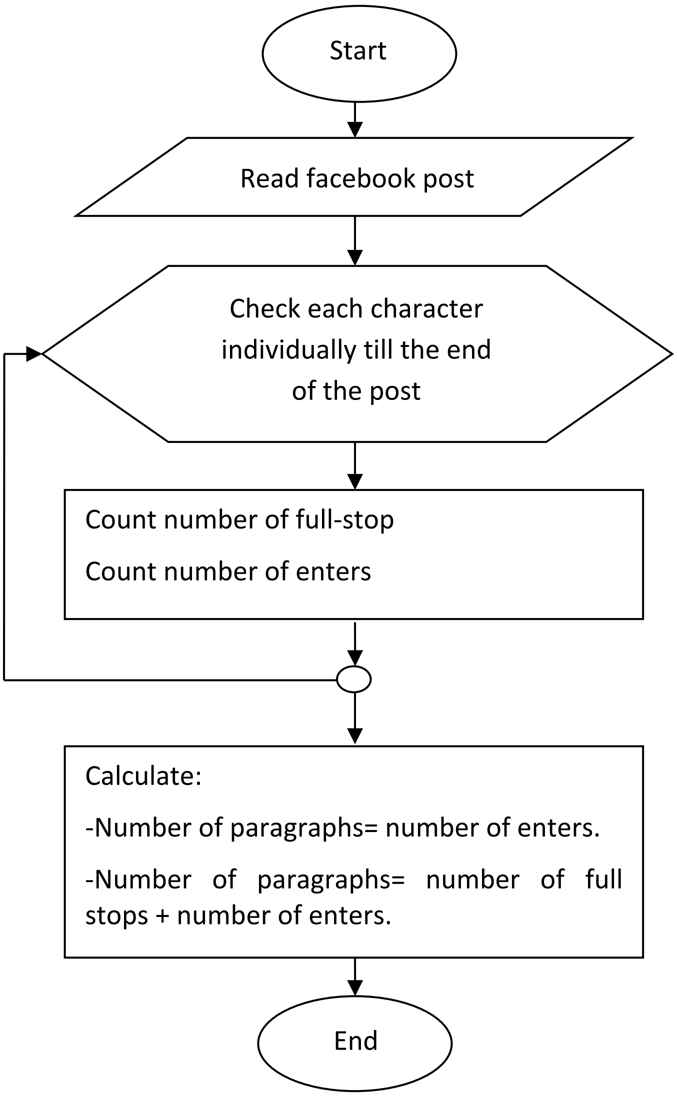
Start
Read facebook post
Check each character individually till the end of the post
Count number of full-stop
Count number of enters
Calculate:
-Number of paragraphs= number of enters.
-Number of paragraphs= number of full stops + number of enters.
End

Figure (2): The flowchart of extracting paragraph-based features

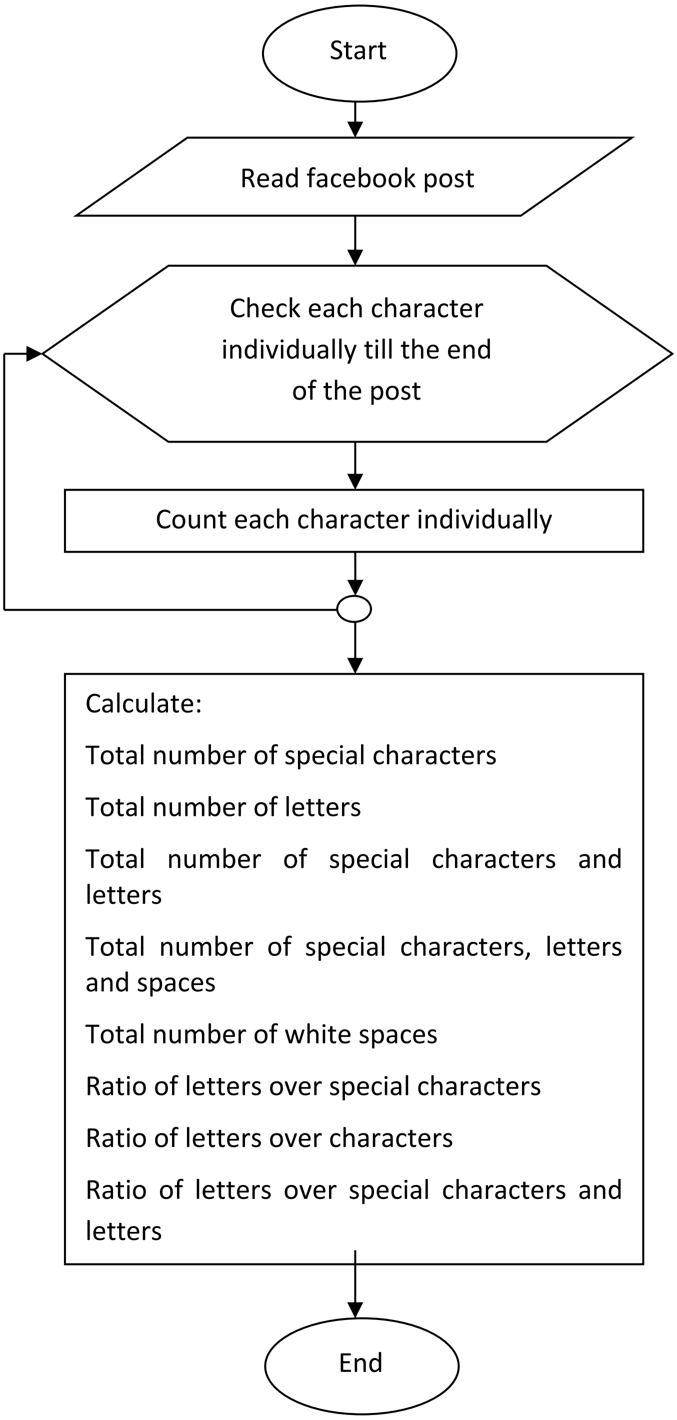
Start
Read facebook post
Check each character individually till the end of the post
Count each character individually
Calculate:
Total number of special characters
Total number of letters
Total number of special characters and letters
Total number of special characters, letters and spaces
Total number of white spaces
Ratio of letters over special characters
Ratio of letters over characters
Ratio of letters over special characters and letters
End

Figure (3): The flowchart of extracting character-based features

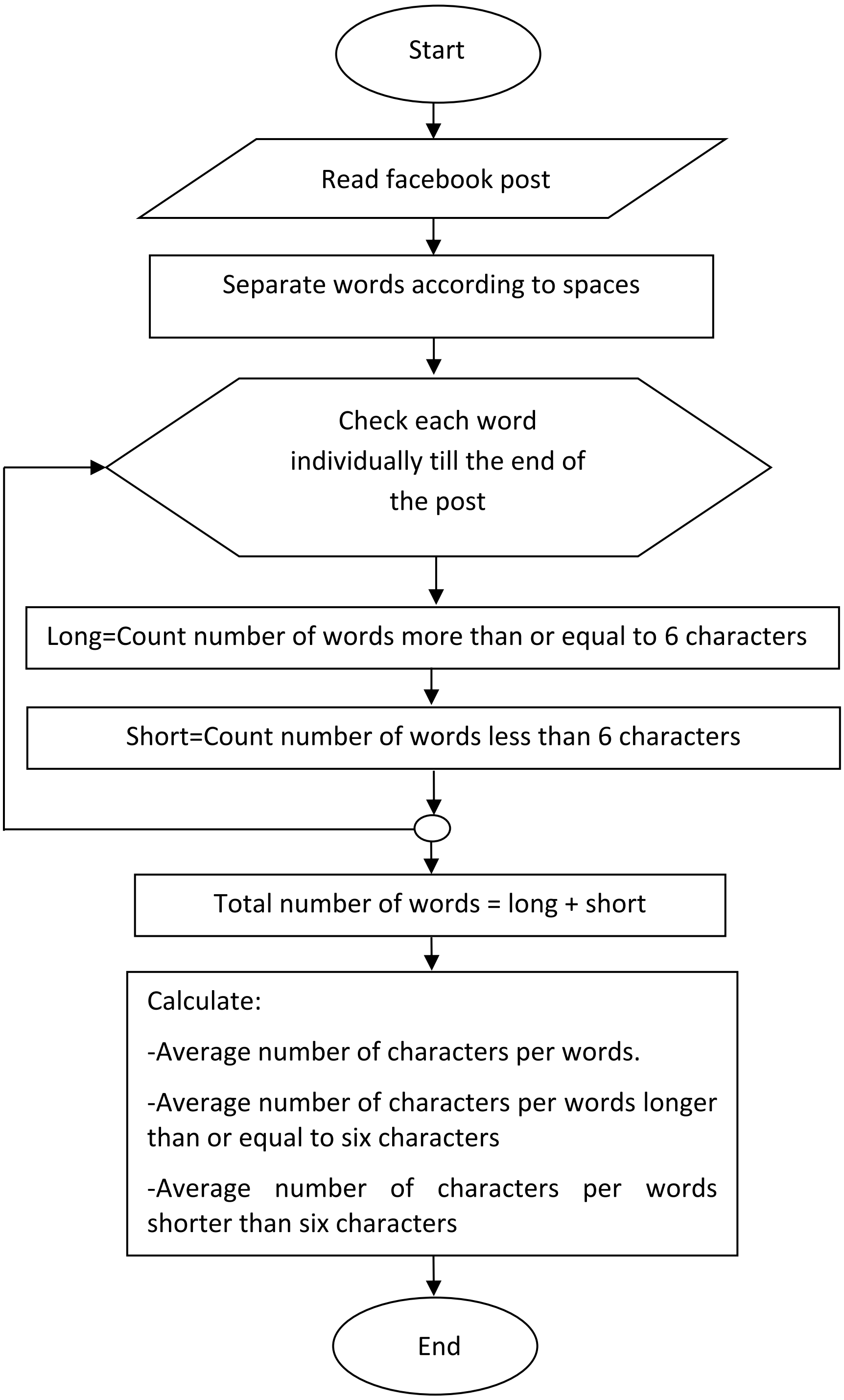
Start
Read facebook post
Separate words according to spaces
Check each word individually till the end of the post
Long=Count number of words more than or equal to 6 characters
Short=Count number of words less than 6 characters
Total number of words = long + short
Calculate:
-Average number of characters per words.
-Average number of characters per words longer than or equal to six characters
-Average number of characters per words shorter than six characters
End

Figure (4): The flowchart of extracting word-based features

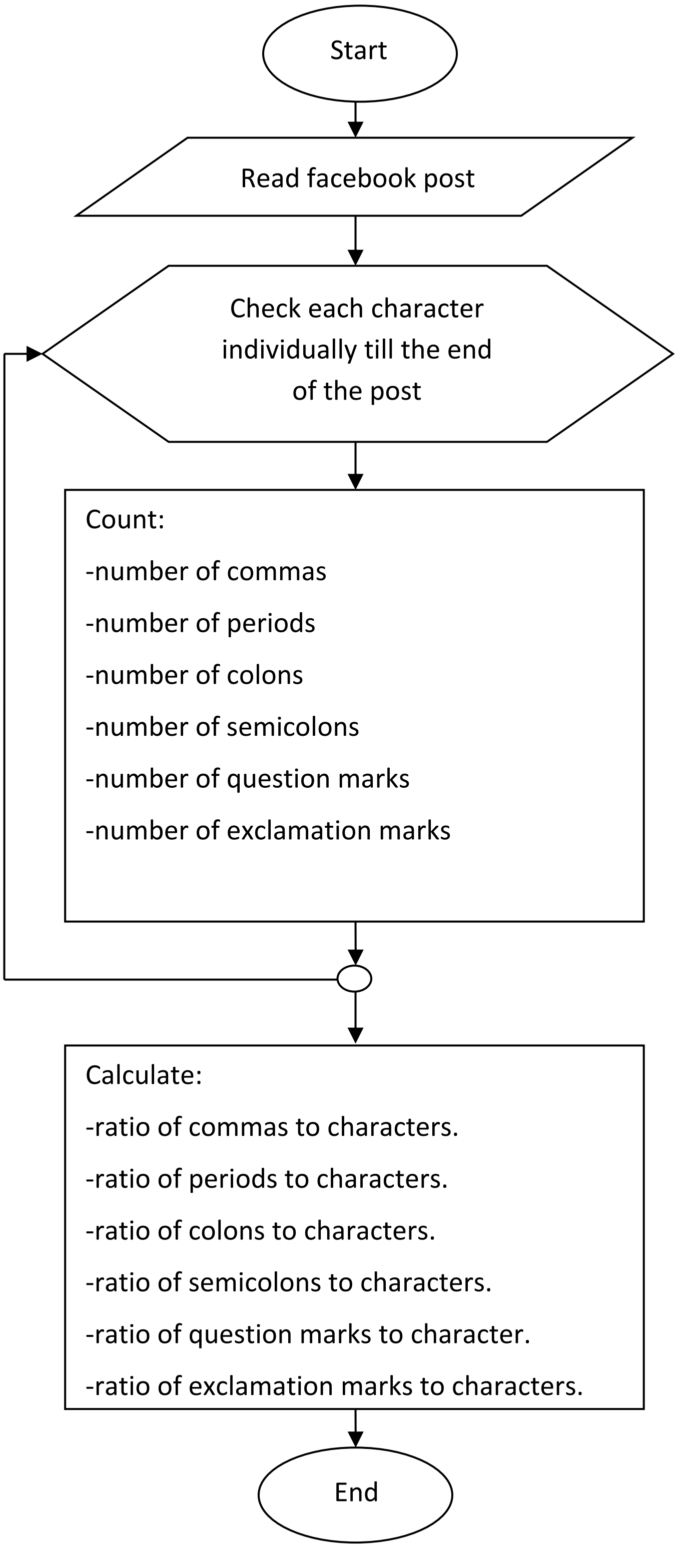
Start
Read facebook post
Check each character individually till the end of the post
Count:
-number of commas
-number of periods
-number of colons
-number of semicolons
-number of question marks
-number of exclamation marks
Calculate:
-ratio of commas to characters.
-ratio of periods to characters.
-ratio of colons to characters.
-ratio of semicolons to characters.
-ratio of question marks to character.
-ratio of exclamation marks to characters.
End

Figure (5): The flowchart of extracting syntactic-based features

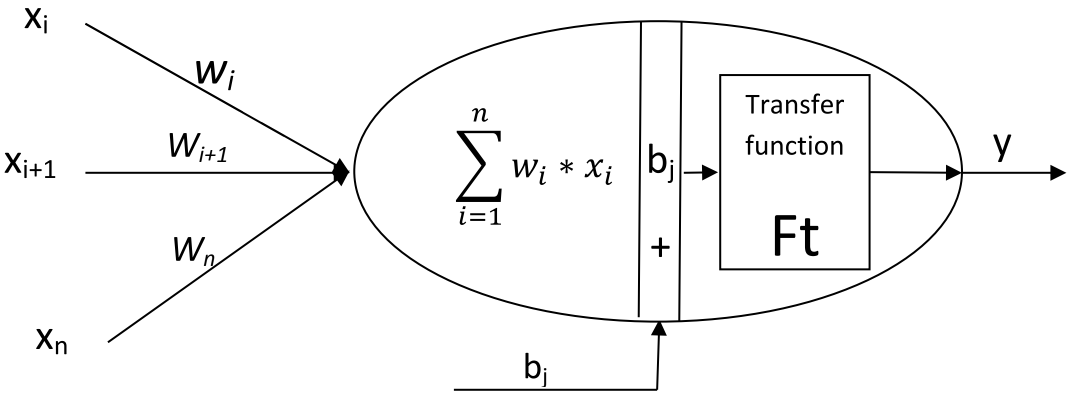

Figure (6): A simple perceptron

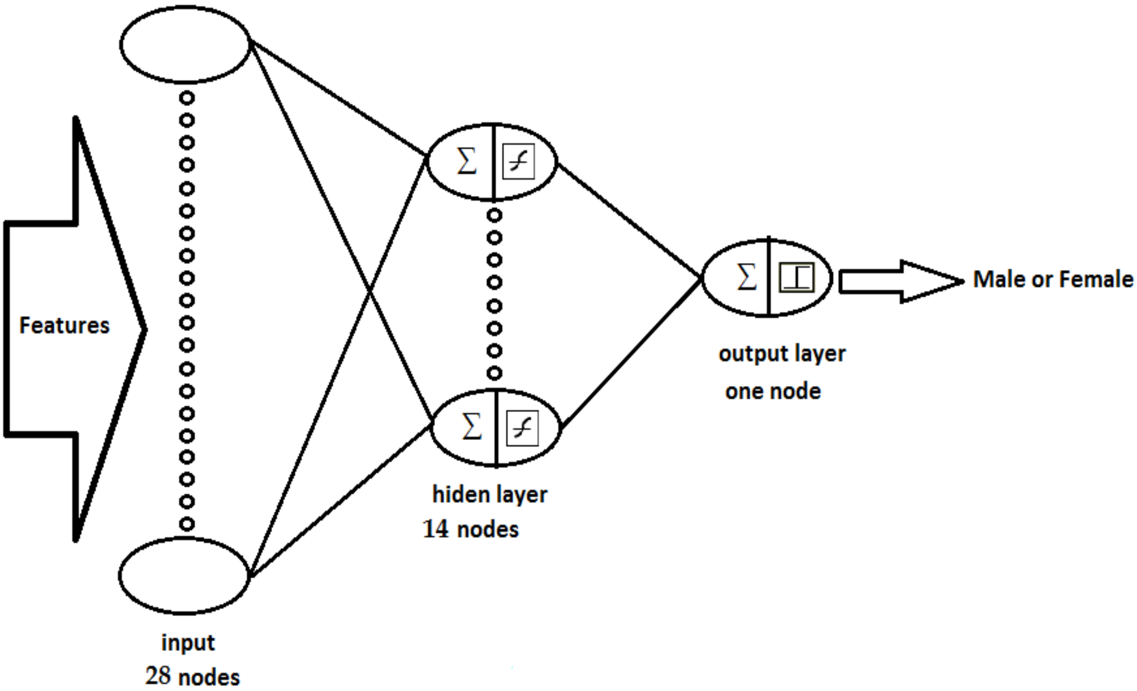

Figure (7): Feed-forward neural network model

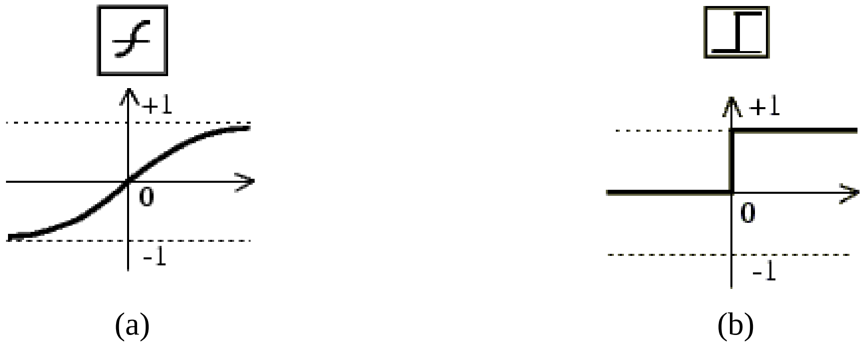

Figure (8): (a) hyperbolic tangent function. (b) hard limit function