International Journal of Future Generation Communication and Networking
Vol. 13, No. 4, (2020), pp. 822 – 827

A Survey On Current Techniques In Authorship Identification

Name:
Rajalekshmy K.D.,

Research Scholar, Department Of Computer Science, Karpagam Academy Of Higher Education

Reg No: 18jul/Cs/Ph.D.Pt/013

Guide Name:

Dr. S. Sheeja,
Professor, Department Of Computer Science, Karpagam Academy Of Higher Education

Article No: Rs/Cs/463/13.01.2020
Mobile No: 9847406409
Mail Id: Rajidevaraj16@Gmail.Com

Abstract

Authorship Identification or Authorship Attribution is agrowing research area which is concerned with the identification of the real author of a disputed anonymous document from the characteristics of that document. This problem has a long history starting from the 19th century. But it has shown tremendous growth recently by the widespread use of social media like twitter, facebook, instagram etc...These electronic texts use very concise messages with minimum number of words and emotions and are the hosting ground for various online crimes. Classical text classifiers use many linguistic features. There are various studies which use N-grams with high dimensionality for classification purposes. This paper discusses the word2vec representation for generatingword embeddings which preserve the semantic relationship between the words.

Keywords—*Authorship Identification, Performance, Framework, CBOW, Skip-gram, Word embedding, word2vec, character n-grams, word n-grams, Recurrent neural network, Convolutional Neural Networks*

I. INTRODUCTION

Authorship Identification is one of the oldest as well as newest research areas in information retrieval. The developments of modern computers, large corpora, and modern statistical and machine learning techniques have made it possible to solve problems algorithmically. There are mainly 3 problems in Authorship Attribution(AA).

(i) Closed AA problem: Given a sample of text and a set of authors, determine that who has written that document.

(ii) Open AA problem: Given a document, determine that who has written that document.

(iii) Stylometry or Profiling: Determine the sociolinguistic properties of the author, such as the age, sex, educational and cultural background etc... It also determines whether a single author or multiple authors are involved in writing a sample of text.

II THEORETICAL BACKGROUND

Theoretical background of AA reveals the fact there is a unique style of language is used by each author which can be called as author's fingerprint can be visible in their writings.[1] Every person has a unique style and experience in learning a language, their style of writing will also be different in very minute aspects and by incorporating multiple features makes AA more reliable.

ISSN: 2233-7857 IJFGCN

III HISTORICAL BACKGROUND

Earlier studies showed length of words can be used as the characteristic feature of each writer. Longer words are used by authors with large vocabularies. But studies have shown that average word length is varying even for a single author and cannot be considered as a prominent feature for identifying authors.[1] Later studies have revealed that an author's writingstyle isrepresentedby calculatingsummary statistical measures of either onefeature or a list of features extracted from a given document. This statistic varies consistently for different authors. A relative vocabulary overlap proposed byUleshows the degree to which two texts are drawn from the same vocabulary. But that method had several drawbacks even though it is sensitive to the differences that summary of statistics gets hide. It has increased theeffort required for analysis by computing the differences on each pair of documents and the topic of the document is dominated over authorship. Later studies have proposed an approach which focus on synonym pairs and authors make a consistent choice of one over the other. But it has become hard to find enough synonym pairs.

Mosteller and Wallace(1964)have focused on function words like prepositions, conjunctions and article which carry little meaning by themselves but preserve the semantic or syntactic relationship with other words. Since, they are topic independent contribute most to the identity of an author. They have analyzed distribution of 30 function words which have been extracted from contents and known to be one of the popular and largely used statistical analysis technique for authorship identification. [1] This method is the best-knownand has success in stylometry. Since, the function words determine the syntax of a sentence; they are called as syntactic features. Other syntactic features include POS and punctuation usage. Later studies have revealed that syntactic features can be used as reliable features in researches involvingauthorship identification.

Later it has been discovered that structural features dominate over syntactic and lexical features. Writers organize articles in their own style by keeping the length of a paragraph, indentation, usage of capital letters and other special characters consistently throughout their articles. Structural features are more prominent in online documents because of its flexible structure and have less content information.

The performance of authorship identification can be improved by using content-specific features which are the key-word-based features used consistently and very much related to the contents of those articles. These features are better than lexical features but not as good as function-words.

In general, a set of selected features and analytical techniques together can be used to evaluate. the performance of authorship identification.

In earlier days, statistical univariate analytical methods were used. The most popular methods were based on histograms, Naïve Bayes Classifiers and cumulative sum statistic. But these methods were greatly discarded due to their instability on multiple topics and their disability to deal with more than two features. Hence multivariate methods were introduced. The first popular multivariate method was Principal Component Analysis (PCA). Later factor analysis, discriminant analysis and cluster analysis were introduced. All this method got wider acceptance due to their good results and hence multivariate approaches were proved to be effective technique in authorship analysis.

Later it has been proved that machine learning techniques are more powerful and reliable in authorship identification than the statistical techniques. Machine learning techniques show higher accuracy, can deal with large number of features, tolerant to noise and non-linear interactions among the features. They require very few assumptions for the mathematical models. Several machine learning techniques such as multilayer perceptron, Radial Basis Function (RBF) network, simple Markov Chain model, Support Vector Machines (SVM) employed by different researchers.

In addition to the features selected, techniques used there are parameters which also contribute to the performance of authorship identification. These parameters include number of authors to be identified, training set size used for training the classification model etc...

It is also very important to study authorship identification involving multiple languages because of the world wide use of Internet. Features and feature extraction techniques vary according to the linguistic characteristics of the languages.

One challenge in authorship identification is to attribute an online text of limited length to an author. It is very difficult to identify the characteristics of a writer from short online messages. At the same

ISSN: 2233-7857 IJFGCN

time online messages have some characteristics such as layout features of the structure, usage of unusual content markers which are very unusual and sub stylistic features which can be helpful in creating the feature set.

IV RELATED WORK

A framework was proposed by Rong Zheng et.al (2006) for authorship identification which consisted of four steps: (i) Message collection (ii) Feature Extraction (iii) Model Generation and (iv) Author Identification. They have used newsgroup messages in English and Chinese languages as their dataset for testing the framework. They have extracted lexical, syntactic, structural and content- specific features and added these features incrementally to study their effects. They have employed three different techniques for model generations: C4.5, Neural Network and SVM.SVM and Neural Networkhave showed better performance compared to the model generated by C4.5. But their study could not identify the minimal feature set with good performance. [2]

EfstathiosStamatatos (2009) presented profile-based, instance-based and hybrid approaches for authorship identification. He also has discussed the issues in authorship identification which would attract the future work in this field. [3]

Robert Layton et.al (2010)have showed that SCAP (Source Code Authorship Profile) methodology is found to be effective in authorship identification on short twitter messages with message length less than 140 characters or less. SCAP methodology has used profile-based approach where each author's documents are concatenated to form a single file of documents and calculated the most frequently occurring n-grams for the combined document. A list is formed with top most n-grams. This list is termed as Simplified Profile of the author. Simplified Profile is formed for each testing document and calculated the similarity measure based on SPI(Simplified Profile Intersection) which is an effective distance metric more robust than relative distance.Thus, it is proved more effectively that the document belongs to the author with highest similarity value.A threshold of 120 tweets per author must be chosen;later addition of more tweets would not give significant increase in accuracy. This method also suggests that accuracy can be improved significantly by considering authorship analysis and analysis of the group of users who are in close communication with the given user.[4]

Roy Schwartz et.al (2013) have used Twitter as their experimental test bed. They have introduced a new concept called k-signatures. These are the features which appear in at least k% of the training samples of an author, but at the same time it is not appearing in the training samples of any other author. As the value of k increases, unique style of the author increases. They have used SVM classifier with character n-grams and word n-grams features. They have presented a new feature called flexible pattern which are a generalization of word n-grams and capable of finding fine-grained differences between authors' styles. They improved the classification results greatly. [5]

Mudit Bhargava et.al (2013) have discussed authorship attribution as two stage process: stylometric information extraction and classification. They have used twitter dataset and this dataset is not biased towards specific content, user or geographic area as the experimental test bed. The extracted stylometric features have been used in SVM classifier with RBF kernel and show good performance and improvement. [6]

Siwei Lai et.al (2015) have applied recurrent structure which can be used for extracting the contextual information while learning word representations and employed a max-pooling layer which automatically capture key features in texts. They have utilized the advantages of both Recurrent Neural Networks and convolutional Neural Networks. [7]

Armin Heonen (2017) has used word embeddings for the process of authorship identification. Word embeddings preserve the semantic relationship. He has used German and English as the corpora and extracted word embeddings for each corpus separately. He has compared sets of similar words and aggregated and computed average values of similarities per text pair. The extent of similarity measuresis then used to distinguish the author of an anonymous document. [7]

NacerEddineBenzebouchi et.al (2018)have proposed word embeddings for feature extraction and used Convolutional Neural networks, Recurrent- Convolutional Neural networks and Support vector Machines as classifiers. The final decision is obtained by using voting method. [8]

ISSN: 2233-7857 IJFGCN

V DISCUSSIONS

Machine Learning and Deep Learning algorithms cannot access text directly and need some numerical representation of the data so that algorithms can process the data. Simple machine learning algorithms use TF-IDF which does not preserve any relationship with words. Word embeddings, which are the neural representation of the words in a document map all words present in a document to a vector space of a specified dimension. Word2vec is a popular method for generating word embeddings. This converts word into vector and with vectors' multiple operations can be performed such as addition, subtraction, compute the angular distance and these operations are used to preserve the relationship among the words.

Word2vec is two layered neural networks generate word embeddings for a given document corpus. It preserves the semantic relationship between words, deals with addition of new words in vocabulary and shows better results in deep learning applications. The main objective of word2vec representation is to produce similar embeddings for words that occur in similar contexts.

The vectors are generated from words using 2 models: Continuous Bag of Words (CBOW) and Skip-gram models. CBOW predicts a missing word given a window of context words or word sequences whereas Skip-gram predicts the context words given a single word.

CBOW model (figure 1) comprises of three identifiable layers: (i) Input layer (ii) Hidden layer (iii)Output layer. Number of input layer nodes and output layer nodes depend on the vocabulary size. Number of hidden layer nodes represents the dimensionality. Dimensionality varies from one to vocabulary size.

While training the model, forward pass is applied on training examples. Then check for errors, if there is any error, apply the backward pass. The main purpose of back propagation is to correct the weight of neurons or optimize the weight of neurons. Repeat this entire process until optimal weights of all neurons are got. After getting the optimal weight of all neurons, the trained CBOW model can be used to predict the next word for the given sequence.

Training the model consists of two phases :(i) Forward Propagation and (ii) Backward Propagation. During the forward propagation, weights of neurons are calculated from the input layer to the neurons present in the hidden layer using summation. Weight calculation from the hidden layer to output layer is done by using a function called Association Function along with calculation of Softmax Function. The association part is the sum of weights of all neurons connected to the output layer and the associated hidden layer weights. The softmax output for the word 'wj'in connection with the given input context words 'wi' can be given by the conditional probability, in terms of exponentials. The main objective of training is to increase the conditional probability in finding the actual output of word which give the input context words.

During the back-propagation weights of the phases are updated from the hidden layer to the output layer by updating the weights of neurons. Then the weights are updated for the neurons present in the input layer to all the ways to the hidden layer.

During testing phase, given input words in the model will predict the missing word which consists of only the forward pass.

Skip-gram model (Figure 2) consists of 3 identifiable layers: (i) Input layer (ii) Hidden layer and (iii) Output layer. The input to the input layer is a single target word and outputs from the output layer are nearest k semantically and logically related to context words with highest probabilities where k is a constant used to represent the window size. The number of input layer nodes is equal to number of distinct words in the vocabulary.

Number hidden layer nodes represent the dimensionality of the vector. Hidden layer in the skip-gram model only performs summation function. The sum of all weighted neurons is given as input to the nodes in the hidden layer.

The output layer has k context windows. The number of nodes in each context window is equal to the vocabulary size (number of input layer nodes). Each context window will give the most probable context word as the output i.e. k outputs.

The entire operation of this model is divided into 2 parts: (i) Forward pass and (ii) Back propagation. Forward pass takes the inputs, calculates the weight at the hidden layer, generates the outputs and calculates the weights at the output nodes. The nodes in the output layer perform summation and

ISSN: 2233-7857 IJFGCN

activation functions. The activation function is the softmax function. After getting the output if there is any error, apply the back propagation to get the optimal weights of neurons.

The testing phase calculates the outputs automatically from the given inputs. There is no back propagation during this phase.

CBOW is preferred when the size of the corpus is small and performs the training faster.Even though Skip-gram performs the training slower, it works well when the size of the corpus ishuge and involves large dimensions. To increase the accuracy, adequate amount of data may be added to the datasets and increase the dimensions of word vector to preserve more information and increase the window size.

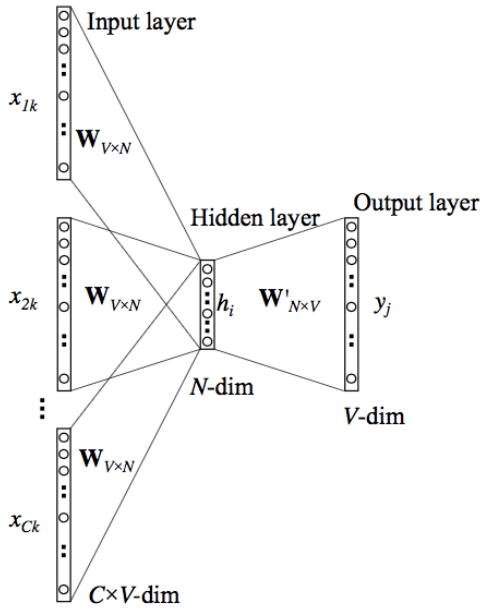

Figure 1: CBOW architecture

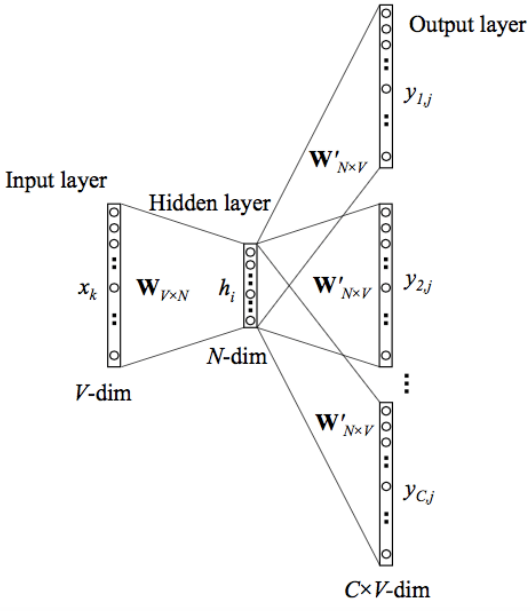

Figure 2: Skip-gram architecture

VII CONCLUSION

This paper discusses the current neural representation of the words, word2vec and two different models CBOW and Skip-gram. A lot of deep learning applications involving in the text have shown

ISSN: 2233-7857 IJFGCN

improvement after using word2vec embeddings as features. This paper also suggests the use of this representation for authorship identification in online messages.

REFERENCES

[1] Patrick Joula, "Foundations and Trends in Information Retrieval", Vol.1, No.3 (2006) 233-334 © 2008.

[2] Rong Zheng ,Jiexun Li, Hsinchun Chen and Zan Huang, " A Framework for Authorship Identification of Online Messages: Writing style Features and Classification Techniques", Jouranal of the American society for Information Scienceand Technology, 57(3): 378-393,2006.

[3] Efstathios Stamatatos, "A Survey of Modern Authorship Attribution Methods", Journal of the American Society for Information Science and Technology, 60(3):538-556,2009.

[4] Robert Layton, Paul Watters, Richard Dazeley, "Authorship Attribution for Twitter in 104 characters or less", Second Cybercrime and Trustworhy Computing Workshop , IEEE 2010.

[5] Roy Schwartz, Oren Tsur, Ari Rappoport, Moshe Koppel , " Authorship Attribution of Micro-Messages", Conference on Empirical Methods in Natural Language Processing, pages 1880-1891, Seattle,Washington, USA ,October 2013.

[6] Mudit Bhargava, Pulkit Mehndiratta and Krishna Asawa , " Stylometric Analysis of Authorship attribution on Twitter", Springer Internatinal Publishing Switzerland 2013.

[7] Siwei Lai, Liheng Xu, Kang Liu, Jun Zhao, " Recurrent convolutional Neural Networks for Text Classification", Association for the Advancement of artificial Intelligence, 2015.

[8] Armin Hoenen, "Using Word Embeddings for Computing Distances Between Texts and for Authorship Attribution" , Springer International Publishing AG 2017.

[9] Nacer Eddine Benzebouchi, Nabiha Azizi, Monther Aldwairi, Nadir Farah, "Multi-Classifier System for Authoship Verification task using Word Embeddings", IEEE 2018.

ISSN: 2233-7857 IJFGCN