INTERNATIONAL JOURNAL OF SOFT COMPUTING

MEDWELL PUBLICATIONS

Text Document Clustering using Hashing Deep Learning Method

Nahrain A. Swidan, Shawkat K. Guirguis and Omar G. Abood
Department of Information Technology, Institute of Graduate Studies Research, Alexandria University, Alexandria, Egypt

Key words: Web news mining, deep learning, LSTM, hash, geolocation

Corresponding Author:
Nahrain A. Swidan
Department of Information Technology, Institute of Graduate Studies Research, Alexandria University, Alexandria, Egypt

Page No.: 44-52
Volume: 14, Issue 2, 2019
ISSN: 1816-9503
International Journal of Soft Computing

Abstract: Web mining is the method of analyzing an grouping of behavioral, statistic, way of life, value-based, web and geographic data for the personalization of offers to online shoppers in genuine time. The goal of this study is to build an effective model for the use of hybrid data clustering and classification technology to evaluate online news data. Assess the best way to use site news information algorithms and assess the reliability of the online news databases use tools and techniques for data mining. A well-known platform to share information among online users is a web-based application. However, nowadays, it is the most challenge to handle gigantic data or enormous information such as web news or web-based promoting by users. On the other side, web applications are the most readily available medium for consumers to access up-to-date information. Such apps also need tremendous space, time and drain the battery power of the mobile devices of the users. One solution to mitigate these challenges is therefore, to extract or extract certain information on the basis of certain characteristics. In contrast, the attributes are the actions or the information collected from different sources by the consumer. This essay attempts to design and implement a web app to extract information on geolocation and space and provides a comparative study on three specific mining techniques.

INTRODUCTION

A significant number of millions of people use the online life day by day. The data is added, edited and read on the web. This is why the world wide web can be viewed as the most immensely colossal database in the world. Data mining specialists have dedicated their careers to better understanding and draw conclusions from the comprehensive information processing by focusing on techniques and innovations in the intersection of database management, statistics and machine learning. This incredible database is a great base for studies in data mining. Basically, data mining reveals unknown patterns in a very big quantities of data. They call it web data mining or web mining if data mining methods are used on web data. Originally, web mining was defined using two distinct methods. Initially, web mining was defined with two distinct methods. The first was a procedure driven view which characterized web mining as a succession (Tang *et al*., 2008).

Firstly, the mining process is a data-centric which characterized web mining as far as the kinds of web data that was being utilized in the mining procedure. The second strategy is more acceptable in the research community in the latest days and will be used in this research (Johnson and Gupta, 2012). Lately, the

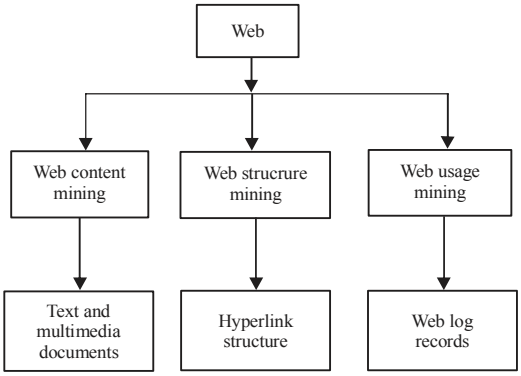

Fig. 1: Web mining taxonomy

subsequent methodology is progressively worthy in the research network and will be utilized in this research. In light of this methodology, characterized web data mining as the entire information mining and related systems that are utilized to consequently find and concentrate data from web reports and administrations. Mining is the biggest mission due to heterogeneity and shortage of shape of web data. Web mining is an application of data mining techniques to find information patterns and relationships from the web data. In order to generate new information, data mining is the method of looking at broad knowledge banks. We might intuitively assume that "mining" information applies to collect new data but that is not the case; rather, data mining is about extrapolating trends and new knowledge from text mining is the process of mining useful information from text documents (Al-Asadi *et al*., 2017). Text mining techniques are used in different types of research domains such as natural language processing, text classification, information retrieval and text clustering. Web data processing is the method of handling a high volume of data. The process of handling/processing data is not easy as explained in previous research. Therefore, researchers utilize web mining, deals with identifying patterns which the user requires. The second phase of web mining is called web content mining which deal with mining of pictures, text and graphs, etc. As with any zone of information, the net comes with a parcel of language. In any case, there are many fundamental terms you wish to understand at the beginning, since, you'll listen these expressions all the time as you examined on. It's simple blend upto blend up these terms in some cases, since, they allude to related but diverse functionalities. In reality, you can see these terms abused in news reports and somewhere else, so, getting them blended up is understandable. Web data processing is the method of handling a high volume of data web mining is divided into three main categories depending on the type of data as web content mining, web structure mining and web usage mining (Hussein and Mousa, 2010; Pandia *et al*., 2011; Siddiqui and Al jahdali, 2013) (Fig. 1).

Web content mining: Web content mining is the process by which the contents of web documents are extracted useful information. Content data reflect the factual collection of a web page that has been designed to communicate to users. It may consist of lists and tables, documents, photographs, audio, video or organized records (Pandia *et al*., 2011; Srivastava *et al*., 2005).

Web structure mining: Our goal is to provide a structured overview of web pages and websites. This illustrates the user-to-web connection. It reveals the inter-document link structure of hyperlinks (Pandia *et al*., 2011; Srivastava *et al*., 2005). It also helps to uncover the file layout used to show the architecture of the websites and the architectures of the websites can be compared (Sharma and Gupta, 2012). At that point, It ought to be partitioned into two bunches each one contains the common sort of auxiliary subtle elements utilized (Srividya *et al*., 2013).

Hyperlinks: In connecting web pages to different places, hyperlinks can be used either on the same website or on the other website. A link is divided into two categories, i.e., hyperlink and hyperlink intra-document. The hyperlink intra-document links many sections of one page whereas the hyperlink inter-document communicates between the two sites.

Document structure: The substance of the net page was organized as tree structures based on distinctive HTML and XML labels (Srivastava *et al*., 2005; Sharma and Gupta, 2012).

Web usagemining: The text document in the form of unstructured data. To extract information on the matching pattern from unstructured data. The keywords and sentences are tracked and then keywords are associated with the message. The technique is very useful if there is a large volume of text. Extracting information makes the unstructured text more ordered. Next, the information is retrieved from derived data and then the missing knowledge is identified using different types of rules. Incorrect software assumptions were discarded in the processing of knowledge (Sharma and Gupta, 2012; Herrouz *et al*., 2013).

Pattern discovery: Date results from the preprocessing stage can be used for discovering patterns.

Topic tracking: It approach tracks the user's records and analyses the user profile. This anticipates user-related documents. Yahoo tracks the topic, the user gives a keyword and the user is notified when anything relates to the keyword. Most sectors can use this strategy. The disciplines of medicine and education are used in all regions. Physicians can easily learn the latest treatments in medical practice. The last source of science-related work is used in learning. For research, this methodology

has the drawback of providing information not relevant to our subject matter when we are looking for our issue (Phyu and Wai, 2019; Jain *et al*., 2017).

Information visualization: Different records were clustered using the procedure. Documents are not grouped according to predefined subjects. It's done on flying. There may be records of different groups. This does not lead to the omission of useful papers from the findings. Users can select the topic of interest using this method (Srivastava *et al*., 2000).

The classification techniques: In web usage mining, classification is widely used. It is a supervised learning technique, in which the models are designed to be classified into data classes. It uses various algorithms that act asclassifier. The classification techniques can also be used for studying the user-client behavior and also interesting patterns can be generated. We can classify the relevant and irrelevant links which are visited by a particular user. This can be identified based on the time spent on a particular web page and also the number of hits (Choi and Yao, 2005). Deals with personalization of various web services. User sessions are separated based on the user's access and then these sessions are accessed from the server weblog. Two new approaches were defined for web mining. The first method is additionally known as "process-centric view" characterizes web mining as a arrangement of tasks and within the moment strategy which is additionally, known as "data-centric view," web mining depends on the sort of information utilized. By Silwattananusarn and Tuamsuk (2012) both temporal pattern extraction and association rule mining are combined to frame the classification framework. This method uses IF-THEN rules which have temporal patterns on the left-hand side and prediction is done on the right-hand side. The prediction is done on temporal patterns and important events. By Sebastiani (2002) classification is done in three phases: the first phase is the training phase which uses labeled records. Second is the test phase which uses unseen labeled records. The final phase is the deployment phase which classifies the unlabeled records. In this study, a method is proposed on the net structure mining approach based on a profound learning algorithm. The profound learning demonstrate include the commitment within the proposed approach, three fundamental highlights are considered for amassing the internet substance. Profound learning calculation input will be the above-listed highlight that delivered a few show parameters. The main contributions of the proposed model:

- A deep learning approach is utilized to build and extract the knowledge from web contents
- Three characteristic based highlights are utilized for recognizing critical blocks
- The features contains concept highlight, title highlight and arrange include

Web page classification: Based on the number of classes, a classification issue can be isolated into a double classification in which occurrences ought to have a place to one of two classes and into a different lesson classification at which more than one course is characterized. When as it were one name is doled out to an occasion, the classification issue is characterized as single-label classification. But on the off chance that more than one course is doled out to an occasion, the classification is at that point alluded to as multilabel one. We are able to isolate web page classification into flat and progressive classification where categories are parallel within the previous and organized in a progressive tree structure within the last mentioned in which each category may have a few subcategories.

There are many applications of web page classification and some of them are web content filtering, ontology annotation, helped web relevant publicizing and information base development, building, keeping up or extending web registries (web pecking orders), making a difference replying frameworks models to extend the quality of comes about of look, building an proficient system that's based on crawlers or vertical (domain-specific) look motors, moving forward quality of look comes about.

Current solutions: Due toweb is the one of the fastest growing area in the research is web data mining, here listed the recent and most related studies have been produced:

Silwattananusarn and Tuamsuk (2012) displayed the concepts of web mining, other than they given an diagram of web mining strategies and after that they displayed an outline of distinctive sorts of web substance mining devices and conclude with the calculations. In the long run, they surveyed exploratory mining tools and strategies to mine the net substance on the web. The study, recommended by examination and hypothetical audit the advancement of web mining calculations. The parallelization handle of a huge volume of web information mining forms can move forward execution within the future. The parallelization prepare is the suggestion for the long run as the internet information is ceaselessly developing at quick speed.

Lou and Zhang (20107) focused on different techniques, approaches and variety of the research which are helpful and patent as the important field of data mining technologies. Eventually, they gave an overall thought about the data mining techniques which can be used on various server log files to find the most frequent patterns. Data mining techniques can be used to find user behavior over the web.

Allahyari *et al.* (2017) gave method/analysis: The study classified news data into four predefined classes (business, entertainment, sports and technology). They used the WEKA data mining tool for text classification. Many classification studies were applied to the news

dataset. Anothersolid study has been done onthese algorithms to check accuracy, errors, time, errors and ROC to predict the best algorithm for news dataset classification.

Palma and Zhou (2017) presented a proposed data selection framework for the k-means algorithm to get high precision clusters from the data collection with respect to traditional k-means-type algorithms in three respects. First, in the cluster learning process, they took the changed value of the cluster's Bregman information which is generated by merging one data item into the potential clusters, as the measure of data item's clustering tendency. Second, only data items with strong clustering tendencies, that was the changed value of cluster's Bregman Information was less than the predefined radius were selected to learn the cluster patterns while the remaining data points were ignored and belong to no cluster. The clustering is non-exhaustive. Third, the radius of the clusters can be changed in the learning process. It was a dynamic learning framework. Experiments showed the effectiveness of the proposed algorithm based on synthetic, document and image data.

Choi and Yao (2005) described several of the most fundamental text mining tasks and techniques including text preprocessing, clustering and classification. Additionally, they briefly explained text mining in biomedical and health care domains.

Wu and developed a web scraper specialized in forums. They selected the most appropriate method for the task among three different methods used for text extraction are implemented and tested. The methods were word count, text-detection framework and text-to-tag ratio. The handling of link duplicates was also considered and solved by implementing a multi-layer bloom filter. The results indicate that the text-to-tag ratio has the best overall performance and gave the most desirable result in web forums. Thus, this was the selected method to keep on the final version of the web scraper. The subject of text classification was well studied in the compared papers that define all their characteristics and reviewed all relative methods. However, in the case of web page classification, limited review and survey papers are developed.

By Song *et al*. (2005) the researchers presented the automatic web page classification systems and the techniques used to build it. It is starting with characterizing the net page classification and a depiction of two sorts of classifications: genre-based classification and subject-based classification. The preprocessing steps were carried out to create web pages information appropriate for encouraging machine learning and classification forms. Another, they presented strategies to the dimensional lessening reason and examined the state of the craftsmanship classifiers in terms of web page classification. At last, they assessed numerous web page classifiers. Based on the above analysis of the literature about the technology of web mining, through comparing the difference among technologies and analyzing the main contributions in the research area, if want to research the process of web mining technology using its structure.

Preliminaries: Unlike a deep network, a normal neural network cannot reason with the events that have occurred in the past, as each computation layer of this type of network is independent and does not affect each other. Thus, they are "stateless" and cannot learn the information from the past sequences which are their major drawbacks. By Lopez-Sanchez *et al*. (2019) a spam classification method is developed by using a special architecture known as Long Short Term Memory (LSTM). Before using the LSTM for the classification, the text is should be converted to semantic word vectors with the help of Word2 vec, WordNet and ConceptNet. As known. ConceptNet is heaps better and easier to deal with. Then, the logistic regression is implemented to detect the fault and safe URLs which might generate detection models without a manually feature engineering. This architecture outperforms other deep learning models and feature-engineer models. A survey of many frameworks for the categorization of web pages based on their visual content is proposed by Lopez-Sanchez *et al*. (2017) and Yang *et al*. (2017). Also, the problem of over-time learning is addressed, so, the proposed framework can learn to identify new web page categories as new labeled images are provided at test time. This study builds upon the ideas and results presented by Lee *et al*. (2009), Hinton *et al*. (2012) and Mughal (2018) where the researchers explored the applicability of deep learning techniques to the problem of web page classification by in this study, fake websites were identified using deep learning. In the classification process, feed forward neural networks and stacked automatic encoders are used. To detect fraudulent websites, URLs belonging to websites that are punctuated by the internet are collected and analyzed together with malicious websites.

Problem description and formulation: In the web area, the world wide web acts as two sides, one is a user side and another one is an information provider. Both sides are face problems while dealing with web data. So, web usage mining retrieves useful data. In the modern era, the websites are considered as the one-touch sources of all kinds of information needed by an individual. The data stored in the web spaces are numerous and one can refer to any kind of information with the help of websites. Recently, information is extracted from the web using programmed methods because of the need for information. As the extraction process becomes viral, the websites have become sources of redundant information. The duplication becomes a major issue.

Frequently, data-mining develops over three steps: Text preprocessing is a vital a step of any Natural Language Processing (NLP) system, since the characters, words and sentences known at this stage are the

fundamental units passed to all further processing stages, that assure introducing a better inputs ready made, like many NLP common issue have resolved, e.g., part-of-speech taggers and morphological analyzers, through applications such as machine automatic engine of text translation and information retrieval applications. It includes all needed activities to output a well pre-processed text documents.

Exploration: We must first plan the data, delete duplicates or redundant information and restrict our collection to exactly what we can use. The aim is to delete any unnecessary words or characters that are written for human readability but do not contribute in any way to the function of classification or clustering. Herein, some well known terminologies and methods for text cleaning and preprocessing text data sets:

Removal of stop words, the stop words like “and”, “the”, “a”, “if”, etc are common in all English sentences and are not meaningful in deciding the theme of the article, so, these words can be eliminated from the articles. The solution is to remove these words from the texts and documents (Saif *et al*., 2014).

Removal of punctuation symbols, exclude all punctuation marks from the text (Verma *et al*., 2014). Lemmatization, it is the process of detect all different forms of a term in order to consider all of them as one item, e.g., “contains” “consist” and ”including” would be “contained” (Gupta and Malhotra, 2015; Aggarwal, 2018; Gupta and Lehal, 2009; Dhuliawala *et al*., 2016).

Noise removal: All unnecessary characters should be removed such as punctuation and special symbolsor characters as explained on (Pahwa *et al*., 2018). Also, for the social media text dataset, typographical errors are commonly presented in texts and documents (e.g., Facebook, Twitter). Many solutions were introduced to solve this issue in natural language processing NLP (Mawardi *et al*., 2018; Christanti and Naga, 2018; Dziadek *et al*., 2017).

MATERIALS AND METHODS

Methodology of study: Within the proposed show, distinctive word vector models have been utilized to urge the vectors for words as input. The assignment is to classify surveys either positive or negative but numerous machine learning calculations and profound learning engineering are not able to prepare content specifically. In order to deal with this problem, we need to convert our reviews into word vectors which can then be passed to the deep learning architecture. The main purpose of data mining is to discover patterns between large sections of data and convert data to more accurate/executable information. There is a major focus of organizations including news organizations that deal with the integration of this information to achieve their interest. For instance, doing to create good predictions and decisions in different areas. This is done by means of an online exploration tool that uses data mining algorithms and algorithms to extract information and knowledge directly from the web (Johnson and Gupta, 2012). Every sentence has almost 80-85 parameters in it. But all the parameters are not required for mining purposes. So, feature selection is one of the important steps of web mining. It is also known as the feature matrix. The collected data must be correct and in the proper format. area. First, the data may be inaccurate. Secondly, the data may be incomplete and unavailable. Thirdly, estimation of assurance about the accuracy of the data is simply not possible and also important words like hashtags from every sentence. It will be easier to classify tweets with the help of hashtags. Because of the ill-posed gradient issue within the improvemen with sign activations, existing deep learning to hash ways got to initial learn continuous representations so generate binary hash codes in a separated binarization step, that suffers from substantial loss of retrieval quality. The data collected from various news sources are not appropriate for experimental work. Data clearance must be preprocessed and converted for exploratory analysis into an appropriate type of data. For the online news mining method and as it is well understood that internet news is delivered by web content, it is therefore, necessary to remove HTML and XML tags or text sounds. In addition, web news could include lengthy news or short news and contains enormous text-noise that negatively impacts the mining process. For example, the following processes may be used to clean and integrate/transform site news documents.

Clean up HTML and XML labels online news file. Removing words which are added noise into the web news documents such as “is” “or”, “the” and etc. Since, these words exist in all the documents within different categories. Removal of terms applied to internet media reports like “was” “and” “the” and so on. Since, all records in the different categories include such terms.

To transform the content words into a single case, e.g., to transform all the capital letters and words started in the lower case words, i.e., Google update or Google switch. Target the tokenizing of internet media reports (meaning terms). Since, the techniques for classification operate on descriptive words or tokens.

Removal of a few odd expressions from web media reports. There’s a set number of common terms as the least (for case where the repeating term is as it were overlooked two times and labeled as an exceptional word).

Several records could be cleared of terms after implementing the processes mentioned above by the processes of extraction and washing. It should therefore, be a feature for the empty documents to be published.

Several records could be cleared of terms after implementing the processes mentioned above by the

Table 1: The web document sample of the data-set

Document ID	Text document
1	"fed office weak data caus slow taper"
2	"open-stock fall fed office acceler taper"
3	"ECB focus strong euro drown ECB message keep rate low"
4	"EU week ahead march 1014 bank resolute transpar ukrain"
5	"euro anxiety wane bund top treasuri spain debt ralli"

processes of extraction and washing. It should therefore, be a feature for the empty documents to be published. Once, all these processes are performed and then cleaned documents should be ready for representing knowledgeable data. However, the preprocessed documents still need to transform into an environment that the classification techniques could understand the documents as the input. For web news classification, the tokenized documents could be transformed into the number times that each word appears in the document of the collected dataset. When the transformation function is applied (after the preprocessing functions), then the counts for the words for each document will be as shown in Table 1. This process is to make a table (or matrix), so that, the documents are able to compare with the other documents in the dataset during the classification process. More processes could be applied to the generated matrix/table to produce much more cleaned documents. For example, those words which do not appear more than two times could be removed in the matrix. However, when this process is performed, some of the documents might be empty, therefore, again, the matrix should be checked out from the empty rows or document, if there are, they should be removed from the matrix. Finally, such a learned network or table might be utilized as the input lines of the classification procedure for the reason of mining approaches.

Implementation and analysis of the solution suggested: Different experiments have been carried out to test the use of data and a set of scenarios. This research also shows the results obtained from experiments performed using different settings and criteria to demonstrate the feasibility of the solution suggested.

Preparation and data collection: First, the news aggregator data set is the data set used. "The information system and computer science" "The Computer Learning and Intelligence Institute", "The University of California, Irvine" provided the data set. The data collection is conducted between 10/4/2014 and 10/8/2014. The news is split into four clusters that reflect the web page contents. The data set was initially supplied by the Faculty of Technology, Roma Tre University-Italy's Artificial Intelligence Center (Lee *et al*., 2009). The database comprises of 422937 internet news pages separated into four groups. So, arbitrarily 25.000 internet news pages were chosen as records/text documents for the tests of all types. As shown in Fig. 2, the dataset used for the study is divided. Note: the ' b' label designates

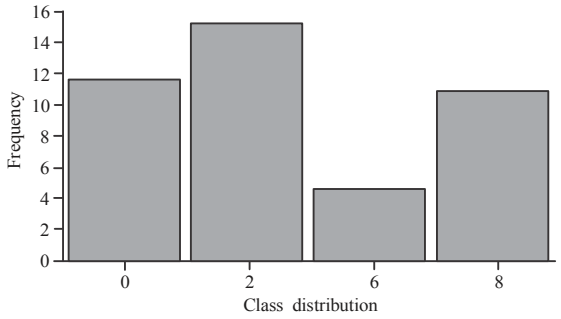

Fig. 2: Distributed web news documents according to class distribution

business, the "e" label refers to entertainment, the "m" label refers to media and the' t' label refers to science and technology. The dataset includes a set of information attributes for various purposes.

The data set is also split into two smaller databases, called learning information and test information sets, for classification/mining purposes. The learning data set is selected randomly and comprises 90-95% of the main dataset's site news, the test data set incorporates 10-5% of the main data set's web news. There are also 22500 reports and media documentation used in the learning testing phase and 2500 records or documents in the test phase. Furthermore, on the two datasets (training and trial datasets), the same preprocessing steps are used which were defined in the preceding subparagraph. The visualized MATLAB 2019 analysis of 5 internet media reports is shown in Fig. 3. As can be seen, the lexicon of the papers is evacuated from sound, accentuation, long and brief terms to speak to the interesting shape of the expressions. By the by, the web news documentation ought to be deciphered into a space that can be utilized to supply the classification procedures as input agreeing to the pre-processing stage. A good domain for web news is the word-speaking, the web news indexing and the frequently used words for every web news/document.

Testbed, functions and scenarios: This simplifies the hash MD5 specifier in an easy-to-digest way. We will first cover the algorithm's general structure. Expansion specifics and compression procedures were individually provided. First, we start with a dataset (sentences). The message is padded and the length of the message is added to the end. It is then split into blocks. Then, the blocks are analyzed one by one. It is necessary to expand and compress every frame. The quality named the current hash status after each compression. The present hash state will be restored as the final hash after the last block is processed. A description of the, for example, procedure:

Sentence:

- "fed office weak data cause weather slow taper"

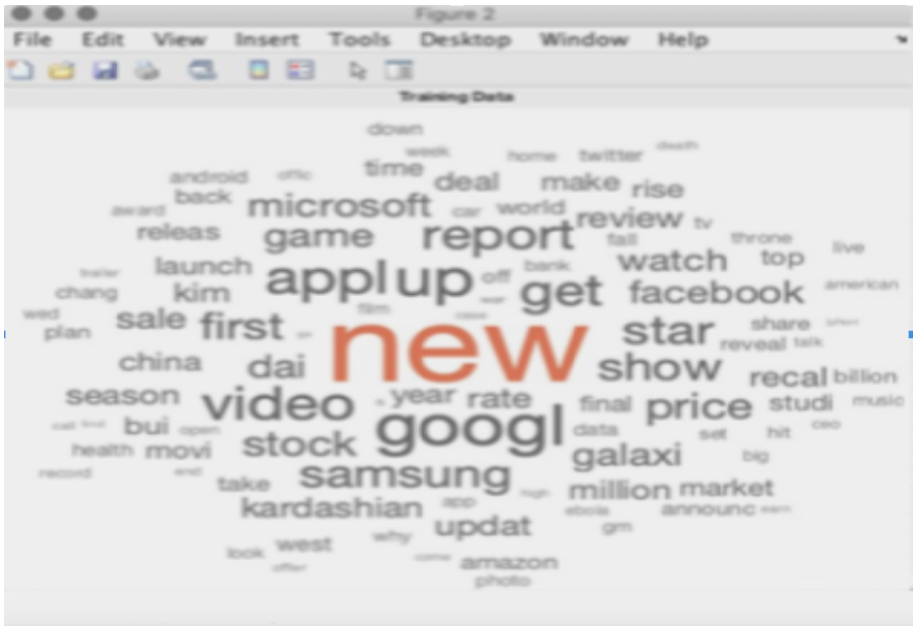

Fig. 3: Training set terms after the pre-processing phase

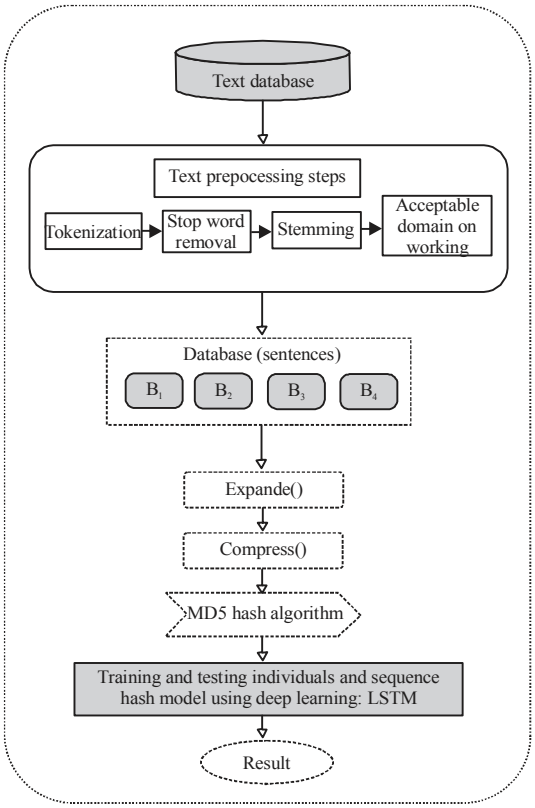

Fig. 4: The proposed model workflow

Tokens:

- "fed", "office", "weak", "data", "cause", "weather", "slow" and "taper"

MD5 hash algorithm output:

"0000000000000000
1111111111
22222222222222222222333333333333333333
44444444444444444444 55555555555555555555
66666666666666666 77777777777777777
888888888888 99999999999999 aaaaaaaaaa
bbbbbbbbbbbbbbbbcccccccccccc
ddddddddddddddd eeeeeeeeeeeeeeeee
fffffffffffffffff"

As shown, hash function maintains the integrity of the items because it chains one block to another. When an item is changed, the hash function is no longer chained to it and the chain is broken. Not every hash function qualifies to chain two items to each other. In the proposed model, a hash function can work with all conditions no needs to meet certain requirements. This is what makes finding an eligible hash more solid (Fig. 4).

RESULTS AND DISCUSSION

Experimental work and results: The proposed model implementation uses MATLAB toolbox for deep learning. Besides usage its classification techniques explicitly Naive Bayes multinomial technique, J48 technique and SMO technique to classify the data and k-means clustering techniques for clustering by Weka tools. In the initial phase, all these three classification techniques are applied to the preprocessed news dataset one after another. And the result of each algorithm is calculated and analyzed. Then compare the result of these three algorithms with each other. The overall analysis of the decision tree, Naive Bayes multinomial and deep learning LSTM technique is shown in Table 2 (Fig. 5).

Table 2: Result for comparisons between the proposed method and other methods

Classification techniques	Accuracy (%)
k-NN technique	85.91
Decision tree	93.29
LSTM	94.05
Proposed method	95.66

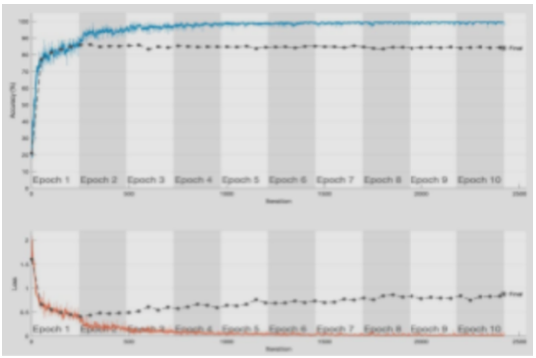

Fig. 5: Deep learning training process

CONCLUSION

The preprocessing of web pages is a very important stage that improves considerably the results of machine learning classifiers and decreases the noisy elements on the web pages. The exploitation of both types of hyperlinks, implicit and explicit one, increases the classification accuracy and enriches the content of the target web page. Deep learning is increasingly chosen in the last 3 years. The advantage of the profound arrange is its capability of learning high-level theoretical highlights continuously. This is often possible due to the passing information learned within the past layers to long-standing time layers. Within the case of web page classification, we outline each web page to one category or numerous categories. This classification plays an vital part in information extraction frameworks as well as look motors, relevant web promoting and others. In this research, we reviewed the existing deep learning algorithms used for web page classification, we produced a literature review and we compared related methods based on some characteristics. For future work, the visual analysis of web pages, the removal of the noisy content and the implicit and explicit links with other pages should be taken into consideration, to have the maximum accuracy possible 95.66%.

RECOMMENDATIONS

It will take less time to merge different classification approaches with further changes for less internet media files. Link the product of the identification to the websites, so that, the manager of the websites can conveniently upload confidential internet news information to the websites. This will be accomplished by saving the online news documents within the content record and after that quickly uploading them to the cloud.

REFERENCES

Aggarwal, C.C., 2018. Machine Learning for Text. Springer, Berlin, Germany, ISBN: 978-3-319-73531-3, Pages: 293.

Al-Asadi, T.A., A.J. Obaid, R. Hidayat and A.A. Ramli, 2017. A survey on web mining techniques and applications. Int. J. Adv. Sci. Eng. Inform. Technol., 7: 1178-1184.

Allahyari, M., S. Pouriyeh, M. Assefi, S. Safaei, E.D. Trippe, J.B. Gutierrez and K. Kochut, 2017. A brief survey of text mining: Classification, clustering and extraction techniques. Comput. Lang., Vol. 1,

Choi, B. and Z. Yao, 2005. Web Page Classification. In: Foundations and Advances in Data Mining, Chu W. and T.Y. Lin (Eds.). Springer, Berlin, Germany, ISBN: 978-3-540-25057-9, pp: 221-274.

Christanti, V.M. and D.S. Naga, 2018. Fast and accurate spelling correction using trie and damerau-levenshtein distance bigram. Telkomnika, 16: 827-833.

Dhuliawala, S., D. Kanojia and P. Bhattacharyya, 2016. Slangnet: A wordnet like resource for English slang. Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16), May 23-28, 2016, Portoroz, Slovenia, pp: 4329-4332.

Dziadek, J., A. Henriksson and M. Duneld, 2017. Improving Terminology Mapping in Clinical Text with Context-Sensitive Spelling Correction. In: Informatics for Health: Connected Citizen-Led Wellness and Population Health, Randell, R., R. Cornet and C. McCowan (Eds.). IOS Press, Amsterdam, Netherlands, ISBN: 978-1-61499-752-8, pp: 241-245.

Gupta, G. and S. Malhotra, 2015. Text documents tokenization for word frequency count using rapid miner (taking resume as an example). Int. J. Comput. Appl., 975: 24-26.

Gupta, V. and G.S. Lehal, 2009. A survey of text mining techniques and applications. J. Emerg. Technol. Web Intell., 1: 60-76.

Herrouz, A., C. Khentout and M. Djoudi, 2013. Overview of web content mining tools. Intl. J. Adv. Res. Comput. Sci. Software Eng., 3: 375-385.

Hinton, G.E., 2012. A Practical Guide to Training Restricted Boltzmann Machines. In: Neural Networks: Tricks of the Trade, Montavon, G., G.B. Orr and K.R. Muller (Eds.). Springer, Berlin, Germany, ISBN: 978-3-642-35288-1, pp: 599-619.

Hussein, M.K. and M.H. Mousa, 2010. An effective web mining algorithm using link analysis. Int. J. Comput. Sci. Inf. Technol. (IJCSIT.), 1: 190-197.

Jain, S., R. Rawat and B. Bhandari, 2017. A survey paper on techniques and applications of web usage mining. Proceedings of the 2017 International Conference on Emerging Trends in Computing and Communication Technologies (ICETCCT'17), November 17-18, 2017, IEEE, Dehradun, India, pp: 1-6.

Johnson, F. and S.K. Gupta, 2012. Web content mining techniques: A survey. Intl. J. Comput. Appl., Vol. 47,

Lee, H., R. Grosse, R. Ranganath and A.Y. Ng, 2009. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. Proceedings of the 26th Annual International Conference on Machine Learning, June 14-18, 2009, ACM, Montreal, Quebec, Canada, ISBN:978-1-60558-516-1, pp: 609-616.

Lopez-Sanchez, D., A.G. Arrieta and J.M. Corchado, 2017. Deep neural networks and transfer learning applied to multimedia web mining. Proceedings of the International Symposium on Distributed Computing and Artificial Intelligence, June 21-23, 2017, Springer, Berlin, Germany, pp: 124-131.

Lopez-Sanchez, D., A.G. Arrieta and J.M. Corchado, 2019. Visual content-based web page categorization with deep transfer learning and metric learning. Neurocomputing, 338: 418-431.

Lou, Z. and C. Zhang, 2017. A data selection framework for K-means algorithm to mine high precision clusters. Proceedings of the 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD'17), July 29-31, 2017, IEEE, Guilin, China, pp: 1651-1657.

Mawardi, V.C., N. Susanto and D.S. Naga, 2018. Spelling correction for text documents in bahasa Indonesia using finite state automata and levinshtein distance method. MATEC. Web Conf., Vol. 164,

Mughal, M.J.H., 2018. Data mining: Web data mining techniques, tools and algorithms: An overview. Int. J. Adv. Comput. Sci. Appl. (IJACSA.), 9: 208-215.

Pahwa, B., S. Taruna and N. Kasliwal, 2018. Sentiment analysis-strategy for text pre-processing. Int. J. Comput. Appl., 180: 15-18.

Palma, M. and S. Zhou, 2017. A web scraper for forums: Navigation and text extraction methods. B.A. Thesis, KTH Royal Institute of Technology, Stockholm, Sweden.

Pandia, M., S.K. Pani, S.K. Padhi, L. Panigrahy and R. Ramakrishna, 2011. A review of trends in research on web mining. Int. J. Instrum. Control Autom. (IJICA.), 1: 37-41.

Phyu, A.P. and K.K. Wai, 2019. Study on web content extraction techniques. Int. J. Trend Sci. Res. Dev. (IJTSRD.), 3: 2235-2238.

Saif, H., M. Fernandez, Y. He and H. Alani, 2014. On stopwords, filtering and data sparsity for sentiment analysis of twitter. Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014) Vol. 5, May 26-31, 2014, Curran Associates, Inc., Reykjavik, Iceland, ISBN:978-1-63266-621-5, pp: 1610-1617.

Sebastiani, F., 2002. Machine learning in automated text categorization. ACM Comput. Surveys, 34: 1-47.

Sharma, A.K. and P.C. Gupta, 2012. Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining. Int. J. Adv. Res. Comput. Eng. Technol., 1: 287-293.

Siddiqui, A.T. and S. Al Jahdali, 2013. Web mining techniques in e-commerce applications. Int. J. Comput. Applic., 69: 39-43.

Silwattananusarn, T. and K. Tuamsuk, 2012. Data mining and its applications for knowledge management: A literature review from 2007 to 2012. J. Data Mining Knowledge Manage. Process, 2: 345-351.

Song, M.H., S.Y. Lim, D.J. Kang and S.J. Lee, 2005. Automatic classification of web pages based on the concept of domain ontology. Proceedings of the 12th Asia-Pacific Software Engineering Conference (APSEC'05), December 15-17, 2005, IEEE, Taipei, Taiwan, pp: 1-7.

Srivastava, J., R. Cooley, M. Deshpande and P.N. Tan, 2000. Web usage mining: Discovery and applications of usage patterns from web data. ACM SIGKDD Explorat., 1: 12-23.

Srivastava, T., P. Desikan and V. Kumar, 2005. Web Mining-Concepts, Applications and Research Directions. In: Foundations and Advances in Data Mining, Chu W. and T.Y. Lin (Eds.). Springer, Berlin, Germany, ISBN: 978-3-540-25057-9, pp: 275-307.

Srividya, M., D. Anandhi and M.I. Ahmed, 2013. Web mining and its categories-a survey. Int. J. Eng. Comput. Sci. (IJECS.), 2: 1338-1345.

Tang, C., C.X. Ling, X. Zhou, N. Cercone and X. Li, 2008. Advanced Data Mining and Applications: 4th International Conference, ADMA. Springer, Berlin, Germany, ISBN: 978-3-540-88192-6, Pages: 759.

Verma, T., R. Renu and D. Gaur, 2014. Tokenization and filtering process in RapidMiner. Int. J. Applied Inf. Syst., 7: 16-18.

Wu, Y.C., 2016. Language independent web news extraction system based on text detection framework. Inf. Sci., 342: 132-149.

Yang, B., X. Fu, N.D. Sidiropoulos and M. Hong, 2017. Towards K-means-friendly spaces: Simultaneous deep learning and clustering. Proceedings of the 34th International Conference on Machine Learning (ICML'17) Vol. 70, August 6-11, 2017, Sydney, Australia, pp: 3861-3870.