Multi-Knowledge Electronic Comprehensive Journal For Education And Science Publications (MECSJ)
ISSUE (29), February (2020)
ISSN: 2616-9185

www.mecsj.com

Text Document Clustering using Hashing Deep Learning Method

Nahrain A. Swidan [1], Shawkat K. Guirguis [2], Omar G. Abood [3], Ahmed S. Hameed[4]
[1 and 4]Department of Information Technology, Institute of Graduate Studies and Research,
Alexandria University, Egypt
[1]Nahrain.adnan@hotmail.com; [2] shawkat_g@yahoo.com; [3]omar.ghazi88@yahoo.com;
[4]ahmed1418sami2012@gmail.com.

Abstract:

Internet exploration is the use of data mining systems to discover patterns of information and relationships from internet data via data mining techniques. The aim of this study is to locate an efficient algorithm for web news mining with analysis of web news data using data clustering and classification procedures based on deep learning, as well as to evaluate the best way to use site news information algorithms compared to other technologies, and to assess the reliability of the internet news databases that are used as tools and techniques for data mining. In this study, we used an effective algorithm (H ash) that is used to collect and classify data for the best classification of Internet news with results close to 95. 66% accuracy.

Keywords: Web News Mining, Deep Learning, LSTM, Hash.

الملخص

استكشاف الإنترنت هو استخدام أنظمة استخراج البيانات لاكتشاف أنماط المعلومات والعلاقات من بيانات الإنترنت عبر تقنيات استخراج البيانات. تهدف هذه الدراسة إلى تحديد خوارزمية فعالة لاستخراج أخبار الويب من خلال تحليل بيانات أخبار الويب باستخدام تجميع البيانات وإجراءات التصنيف بناءً على التعلم العميق، بالإضافة إلى تقييم أفضل طريقة لاستخدام خوارزميات معلومات موقع الأخبار مقارنة بالتقنيات الأخرى وتقييم موثوقية قواعد بيانات أخبار الإنترنت التي يتم استخدامها كأدوات وتقنيات لاستخراج البيانات. في هذه الدراسة ، استخدمنا خوارزمية فعالة **(H ash)** التي تُستخدم لجمع وتصنيف البيانات للحصول على أفضل تصنيف لأخبار الإنترنت مع نتائج دقة تقارب **95.66%**.

الكلمات المفتاحية: اخبار الويب على شبكة الانترنت ، التعلم العميق ، ذاكرةطويلة الامد.

1. 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. Data mining reveals unknown patterns in 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 work (Johnson and Gupta, 2012). Lately, the subsequent methodology has been progressively worthy in the research network and will be utilized in this work. 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 the 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 hear these expressions all the time as you examined on. It's a simple blend up to 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 Aljahdali, 2013)

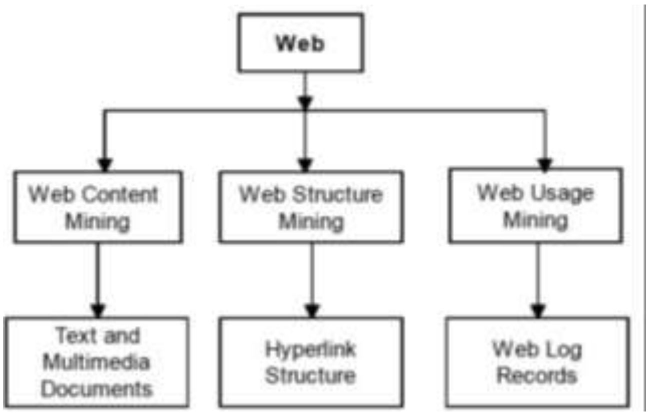


Figure 1. Web Mining Taxonomy

1) WEB CONTENT MINING

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

2) 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 (Srivastava et al., 2005; Pandia et al., 2011). 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).

3) WEB USAGE MINING

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 (Jain et al., 2017; Phyu and Wai, 2019). 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).

A) THE CLASSIFICATION TECHNIQUES

In web use mining, arrangement is broadly utilized. It is a directed learning method, wherein the models are intended to be grouped into information classes. It utilizes different calculations that go about as classifier. The arrangement strategies can likewise be utilized for considering the client customer conduct and furthermore fascinating examples can be produced. We can order the important and insignificant connections which are visited by a specific client. This can be recognized dependent on the time spent on a specific website page and furthermore the quantity of hits (Choi and Yao, 2005). Deals with personalization of different web administrations. Client sessions are isolated depends on the client's entrance and afterward these sessions are gotten to from the server weblog. Two new methodologies were characterized for web mining. The primary technique is moreover known as "process-driven view" portrays web mining as a course of action of undertakings and inside the minute system which is also known as "information driven view," web mining relies upon the kind of data used.

In (Silwattananusarn and Tuamsuk, 2012) both transient example extraction and affiliation rule mining are joined to outline the grouping system. This strategy utilizes IF-THEN principles which have worldly examples on the left-hand side and expectation is done on the right-hand side. The expectation is done on fleeting examples and significant occasions. In Sebastiani, (2002) Sebastiani, the arrangement is done in three stages: the main stage is the preparation stage which uses marked records. Second is the test stage, which utilizes concealed named records.

The last stage is the organization stage, which orders the unlabelled records. Right now, strategy is proposed on the net structure mining approach depends on a significant learning calculation.

The significant learning show incorporates the responsibility inside the proposed approach, three basic features are considered for gathering the web substance. Significant learning computation info will be the above-recorded feature that conveyed a couple of show parameters. The principle commitments 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 contain concept highlight, title highlight and arrange include

A. 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 in a manner of speaking 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 multipliable 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.

2. 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. In Lopez-Sanchez, Arrieta & Corchado, (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 word2vec, WordNet and Concept Net. 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 in (López-Sánchez et al., 2017; 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 paper builds upon the ideas and results presented in (Lee et al., 2009; Hinton, 2012; Mughal, 2018),where the authors 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, feedforward 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.

3. Problem description and formulation

In the web zone, the internet goes about as different sides, one is a client side and another is a data supplier. The two sides are face issues while managing web information. So Web Usage mining recovers valuable information. 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.

1) 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:

a. 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).
b. Removal of Punctuation symbols – Exclude all punctuation marks from the text (Verma et al., 2014).
c. Lemmatization – It is the process of detecting 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 Lehal, 2009; Agarwal and Mittal, 2014; Gupta and Malhotra, 2015; Dhuliawala et al., 2016).
d. Noise Removal: All unnecessary characters should be removed such as punctuation and special symbols or 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 (Dziadek et al., 2017; Christanti and Naga, 2018; Mawardi, et al., 2018).

4. Literature review

Due to web 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:

- Choi and Yao (2005) described several of the most fundamental text mining tasks and techniques including text preprocessing, clustering and classification. Additionally, briefly explained text mining in biomedical and health care domains
- Song et al. (2005) the authors 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. 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,
- Silwattananusarn and Tuamsuk (2012) displayed the concepts of web mining, other than they given a diagram of web mining strategies, and after that they displayed an outline of distinctive sorts of web substance mining devices. In the long run, they surveyed exploratory mining tools and strategies to mine the net substance on the web. The paper 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.

- Wu and Yu-Chieh (2016) developed a web scraper specialized in forums. Chosen the most fitting strategy for the errand among three unique techniques utilized for content extraction are executed and tried. 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 outcomes demonstrated that the Text-to-Tag Ratio has the best generally 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.
- Lou and Zhang (2017) 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) 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. Another solid study has been done on these 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 Bergman 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 Bergman 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.

5. Methodology

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, counting news associations that manage 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 improvement with significant activations, existing deep learning to hash ways got to initial learn continuous representations so generate binary hash codes in a separated banalization step, that suffers from substantial loss of retrieval quality. The data collected from various news sources are not appropriate for experimental work. Information leeway 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 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.

Table 1 The Web Document Sample of the Data- Set

Document ID	Text Document
1	“fed office weak data cause 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”

6. 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 work also shows the results obtained from experiments performed using different settings and criteria to demonstrate the feasibility of the solution suggested.

1) PREPARATION AND DATA COLLECTION

First, the news aggregator data set is the data set used. "The information system and computer science"–"The Computer Learning & 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 Figure 2, the dataset used for the study is divided. Note: the ’ b’ label designates 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 to 95% of the main dataset’s site news; the test data set incorporates 10 to 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 Figure 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.

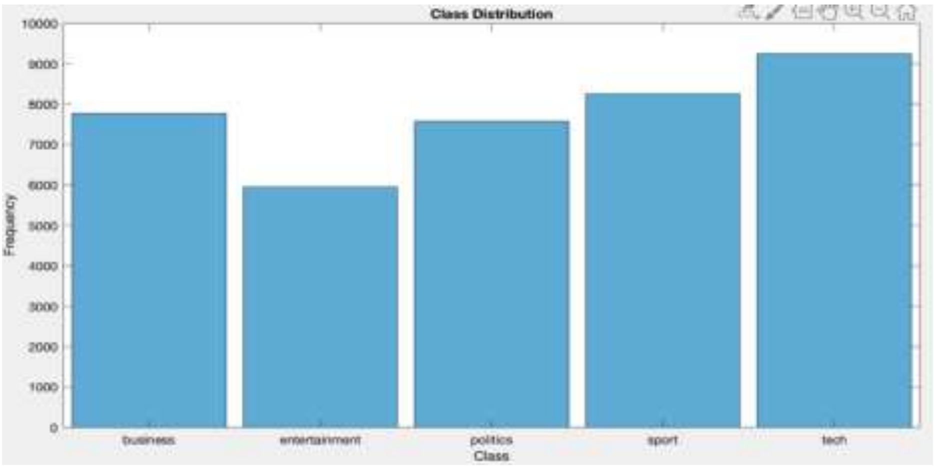


Figure 2. Distributed web news documents according to Class Distribution

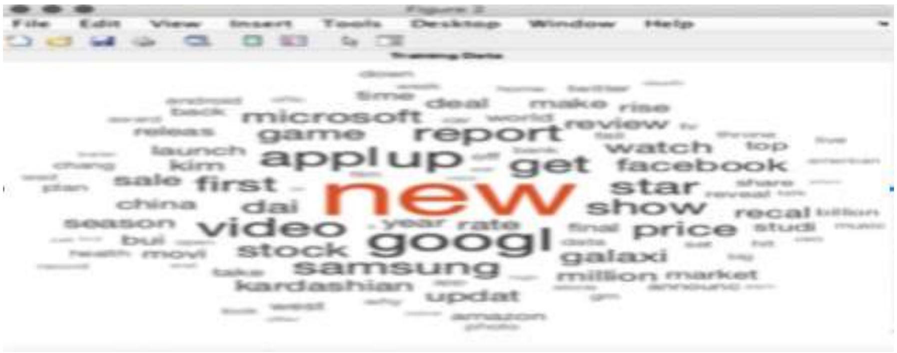


Figure 3. Training Set Terms after the pre-processing phase

2) 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"

Tokens

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

MD5 Hash algorithm OUTPUT

“0000000000000000

1111111111

22222222222222222222 33333333333333333333

44444444444444444444 55555555555555555555

66666666666666666 77777777777777777

888888888888 99999999999999 aaaaaaaaaa bbbbbbbbbbbbbbbbbcccccccccccc

dddddddddddddd eeeeeeeeeeeeeeeeee ffffffffffffffff”

Figure 4. Hash algorithm

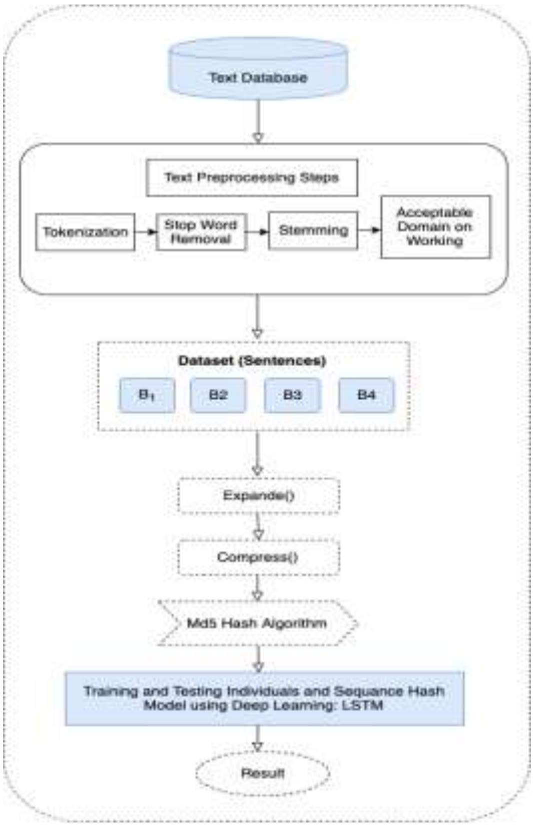


Figure 5. The Proposed Model based on HASH flowchart algorithm

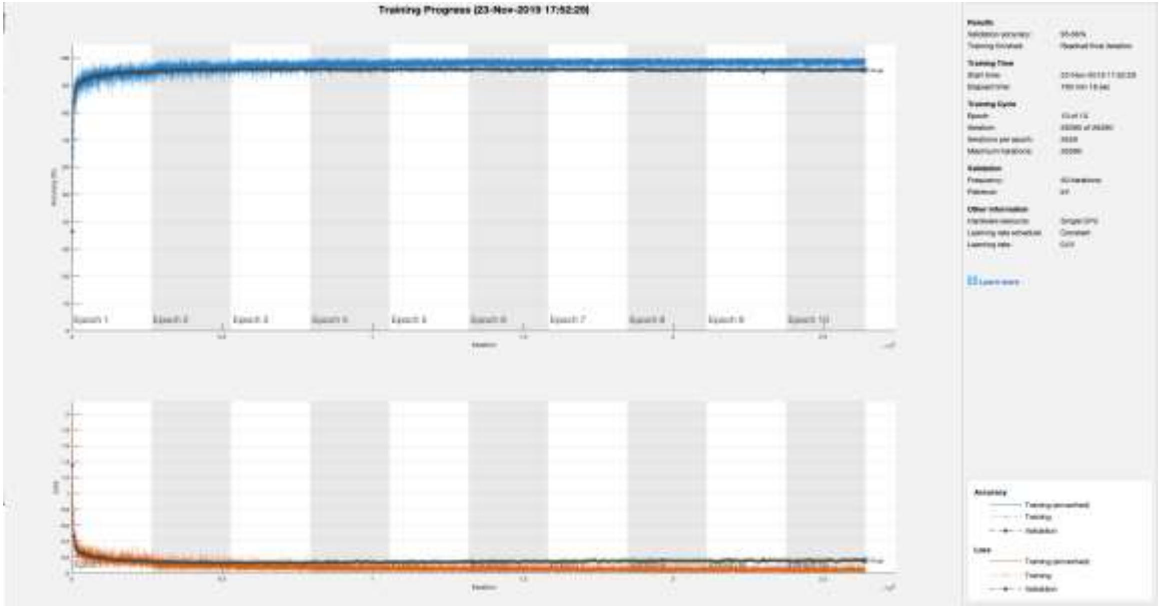
As shown, hash function maintains the integrity of the items because it chains one block to another. When an item is changed, the hash function is no longer chained to it and the chain is broken. Not every hash function qualifies to chain two items to each other. In the proposed model, a hash function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

7. 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.

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%**



Results
Validation accuracy: 95.66%
Training finished: Reached final iteration
Training Time
Start time: 23-Nov-2019 17:52:29
Elapsed time: [illegible]
Training Cycle
Epoch: 10 of 10
Iteration: [illegible]
Iterations per epoch: [illegible]
Maximum iterations: [illegible]
Validation
Frequency: [illegible] iterations
Patience: Inf
Other information
Hardware resource: Single GPU
Learning rate schedule: Constant
Learning rate: [illegible]
Learn more

Figure 6. Deep learning training process

8. 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 three 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 characterization has an indispensable impact in data extraction frameworks as well as look motors, relevant web promoting, and others. In this work 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%.

References

[1] Tang, C., Ling, C. X., Zhou, X., Cercone, N., & Li, X. (Eds.). (2008). *Advanced Data Mining and Applications: 4th International Conference, ADMA 2008, Chengdu, China, October 8-10, 2008, Proceedings*. (Vol. 5139). Springer..

[2] Johnson, F., & Gupta, S. K. (2012). Web content mining techniques: a survey. *International Journal of Computer Applications*, *47*(11).

[3] Al-asadi, T. A., Obaid, A. J., Hidayat, R., & Ramli, A. A. (2017). A survey on web mining techniques and applications. *International Journal on Advanced Science Engineering and Information Technology*, *7*, 1178-1184.

[4] Hussein, M. K., & Mousa, M. H. (2010). An Effective Web Mining Algorithm using Link Analysis. *IJCSIT) International Journal of Computer Science and Information Technologies*, *1*(3), 190-197.

[5] Pandia, M., Pani, S. K., Padhi, S. K., Panigrahy, L., & Ramakrishna, R. (2011). A Review of Trends in Research on Web Mining. *International Journal of Instrumentation, Control & Automation (IJICA)*, *1*(1).

[6] Siddiqui, A. T., & Aljahdali, S. (2013). Web mining techniques in e-commerce applications. *arXiv preprint arXiv:1311.7388*.

[7] Sharma, K., Shrivastava, G., & Kumar, V. (2011, April). Web mining: Today and tomorrow. In *2011 3rd International Conference on Electronics Computer Technology* (Vol. 1, pp. 399-403). IEEE.

[8] Srivastava, T., Desikan, P., & Kumar, V. (2005). Web mining–concepts, applications and research directions. In *Foundations and advances in data mining* (pp. 275-307). Springer, Berlin, Heidelberg.

[9] Sharma, A. K., & Gupta, P. C. (2012). Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, *1*(8).

[10] Srividya, M., Anandhi, D., & Ahmed, M. I. (2013). Web mining and its categories–a survey. *International Journal of Engineering and Computer Science, IJECS*, *2*(4), 1338-1345.

[11] Herrouz, A., Khentout, C., & Djoudi, M. (2013). Overview of web content mining tools. *arXiv preprint arXiv:1307.1024*.

[12] Phyu, A. P., & Wai, K. K. (2019). Study on Web Content Extraction Techniques.

[13] Jain, S., Rawat, R., & Bhandari, B. (2017, November). A survey paper on techniques and applications of web usage mining. In *2017 International Conference on Emerging Trends in Computing and Communication Technologies (ICETCCT)* (pp. 1-6). IEEE.

[14] Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *Acm Sigkdd Explorations Newsletter*, *1*(2), 12-23.

[15] Choi, B., & Yao, Z. (2005). Web page classification. In *Foundations and Advances in Data Mining* (pp. 221-274). Springer, Berlin, Heidelberg.

[16] Silwattananusarn, T., & Tuamsuk, K. (2012). Data mining and its applications for knowledge management: a literature review from 2007 to 2012. *arXiv preprint arXiv:1210.2872*.

[17] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.

[18] Lou, Z., & Zhang, C. (2017, July). A data selection framework for k-means algorithm to mine high precision clusters. In *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)* (pp. 1651-1657). IEEE.

[19] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.

[20] Palma, M., & Zhou, S. (2017). A Web Scraper For Forums: Navigation and text extraction methods.

[21] Wu, Y. C. (2016). Language independent web news extraction system based on text detection framework. *Information Sciences*, *342*, 132-149.

[22] Song, M. H., Lim, S. Y., Kang, D. J., & Lee, S. J. (2005, December). Automatic classification of web pages based on the concept of domain ontology. In *12th Asia-Pacific Software Engineering Conference (APSEC'05)* (pp. 7-pp). IEEE.

[23] Lopez-Sanchez, D., Arrieta, A. G., & Corchado, J. M. (2019). Visual content-based web page categorization with deep transfer learning and metric learning. *Neurocomputing*, *338*, 418-431.

24] López-Sánchez, D., Arrieta, A. G., & Corchado, J. M. (2017, June). Deep neural networks and transfer learning applied to multimedia web mining. In *International Symposium on Distributed Computing and Artificial Intelligence* (pp. 124-131). Springer, Cham.

[25] Yang, B., Fu, X., Sidiropoulos, N. D., & Hong, M. (2017, August). Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (pp. 3861-3870). JMLR. org.

[26] Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009, June). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning* (pp. 609-616).

[27] Hinton, G. E. (2012). A practical guide to training restricted Boltzmann machines. In *Neural networks: Tricks of the trade* (pp. 599-619). Springer, Berlin, Heidelberg.

[28] Mughal, M. J. H. (2018). Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview. *Information Retrieval*, *9*(6).

[29] Saif, H., Fernández, M., He, Y., & Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of twitter.

[30] Verma, T., Renu, R., & Gaur, D. (2014). Tokenization and filtering process in RapidMiner. *International Journal of Applied Information Systems*, *7*(2), 16-18.

[31] Gupta, G., & Malhotra, S. (2015). Text documents tokenization for word frequency count using rapid miner (taking resume as an example). *Int. J. Comput. Appl*, *975*, 8887.

[32] Agarwal, B., & Mittal, N. (2014). Text classification using machine learning methods-a survey. In *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012* (pp. 701-709). Springer, New Delhi.

[33] Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, *1*(1), 60-76.

[34] Dhuliawala, S., Kanojia, D., & Bhattacharyya, P. (2016, May). Slangnet: A wordnet like resource for english slang. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 4329-4332).

[35] Pahwa, B., Taruna, S., & Kasliwal, N. (2018). Sentiment Analysis-Strategy for Text Pre-Processing. *Int. J. Comput. Appl*, *180*, 15-18.

[36] Mawardi, V. C., Susanto, N., & Naga, D. S. (2018). Spelling Correction for Text Documents in Bahasa Indonesia Using Finite State Automata and Levinshtein Distance Method. In *MATEC Web of Conferences* (Vol. 164, p. 01047). EDP Sciences.

[37] Christanti, V. M., & Naga, D. S. (2018). Fast and Accurate Spelling Correction Using Trie and Damerau-levenshtein Distance Bigram. *TELKOMNIKA*, *16*(2), 827-833.

[38] Dziadek, J., Henriksson, A., & Duneld, M. (2017). Improving terminology mapping in clinical text with context-sensitive spelling correction. *Informatics for Health: Connected Citizen-Led Wellness and Population Health*, *235*, 241-245.