BULETINUL INSTITUTULUI POLITEHNIC DIN IAŞI
Publicat de
Universitatea Tehnică „Gheorghe Asachi” din Iaşi
Volumul 66 (70), Numărul 2, 2020
Secţia
ELECTROTEHNICĂ. ENERGETICĂ. ELECTRONICĂ

ANALYSIS OF ONLINE NEWS FOR THE DETECTION OF MISLEADING HEADLINES – A STUDY CASE

BY

TIBERIUS DUMITRIU[*] and CONSTANTIN CROITORU

“Gheorghe Asachi” Technical University of Iași,
Faculty of Automatic Control and Computer Science

Received: May 12, 2020
Accepted for publication: June 12, 2020

Abstract. In recent years media creates a favorable environment for the distribution of fake news. The automatic identification of such type of news is the subject of many debates and attempts to develop computer programs. A particular case of this phenomenon is represented by the attractive headlines that lead the reader to a wrong idea about the content of the article. This paper proposes a framework by which the possible online news with large discrepancy between the title and the content can be identified and signaled to the reader. The k-Nearest Neighbors algorithm and some metrics for measuring similarities were studied to identify if the title is relevant in relation with the content of the article. Java language was used for word processing and to analyze the words frequency in the title and their appearance in the content of the news. The results suggest that these methods may be useful for detecting misleading titles.

Keywords: clickbait detection; fake news; web technologies; artificial intelligence; software applications.

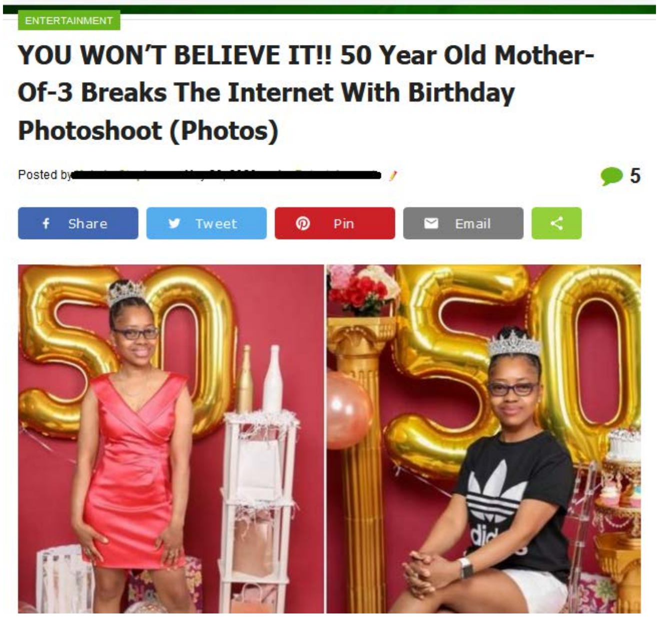
[*] Corresponding author; *e-mail*: tiberius@cs.tuiasi.ro

1. Introduction

The tremendous development of the Internet and the almost exponential increase of access to it facilitated the ways of distributing information. Even though a significant amount of information is spread through television, where the percentage of false or unverified news from several sources is quite small, the news agencies have transferred their way of disseminating information mainly in the online environment. Thus, there are a lot of sources through which information is distributed daily about a wide variety of events. Because the legislation does not prohibit the broadcasting of news with a low degree of veracity, many of the creators of such news distribute it very easily. Being in a great expansion, social networks such as Facebook, Twitter, Instagram and others are perhaps the most important distributors of information to the general public. In this context a favorable environment for the distribution of irrelevant or fake news has been created. In the literature, some articles talk extensively about concrete cases of fake news meant to induce ideas to people, some of them false, other flawed or taken out of context, to determine people to vote with one candidate or the other (Figueira and Oliveira, 2017) or to attract more readers by title even if the content is irrelevant and this affects the credibility and quality of online journalism (Molyneux and Coddington, 2020).

A particular case of this phenomenon is represented by the attractive headlines that lead the reader to a misconception about the content of the article. This is called *clickbait*. Clickbait is a term that describes the strategies to increase the number of views of a page from a site by clicking the main element that lures a reader to access a page or link: the title (Cambridge, 2020). Another possible feature could be that, in order to be able to read the whole article, it will be necessary to access a sequence of pages with very little content, for example a single paragraph, accompanied or not by an image or a video. Besides wasted time and the deviation from the context, the problem of professional ethics and manipulation can also be raised, along with shaping the new generations to the superficial, false and even accepting a more or less refined lie as being something normal and natural.

In order to attract visitors, the titles of web pages may contain expressions such as: “You won't believe it!”, “This will upset you!”, “You have done everything wrong so far!”, “It's outrageous / incredible”. If the titles would be more accurate, then fewer people would access them (Matraguna, 2020). A title that tries to attract visitors for an article containing very little information and a few pictures can be seen in Fig. 1 (Naijaloaded, 2020).



ENTERTAINMENT

YOU WON'T BELIEVE IT!! 50 Year Old Mother-Of-3 Breaks The Internet With Birthday Photoshoot (Photos)

Posted by [redacted]

5

Fig. 1 – A title that wants to impress visitors.

In many papers the effect of clickbait for journalism (Molyneux and Coddington, 2020), economy (Munger, 2020), private or public behavior (Scacco and Muddiman, 2020), emotional arousal (Pengnate, 2019) or political opinion (Allcott and Gentzkow, 2017) were investigated. In this context, some authors have studied and created various applications often using methods in the field of artificial intelligence. Thus, some researchers (Kaur *et al.*, 2020; Agrawal, 2016) used a convolutional neural network to detect the clickbait, while (Xu *et al.*, 2020) used reinforcement learning trying to create an automated clickbait. Clickbait Killer, an extension for Google Chrome browser tries to intercept and block some web pages, but currently removes only a limited list of links from web (Chrome Web Store, 2020).

A reliable application that could notify such titles from the online environment becomes a necessity. Although there are still some software created for this purpose, none of them meets most of the market requirements. One of the issues is related to the time required for news analysis. Being online applications, the user expects a fast feedback, obtained in real time. Other problem to be solved is to decide whether the headline and content are consistent for a particular news story. For this reason it can be considered a classification issue.

In the given context, the main goal of this research is represented by the development of a framework aiming to solve these problems. The news items were classified in two categories (relevant and irrelevant) according to the relevance of the headline to their content. To achieve this goal, some web technologies were used, such as: Java (using the Spring framework) because it facilitates working with word processing, Mongo database for storing data structures and the access speed to them and HTML (Hyper Text Markup Language), CSS (Cascading Style Sheets), JavaScript (with the help of the jQuery framework) were also used. The final application also contains an extension for a Google Chrome browser.

2. Framework Description

In Fig. 2, the framework architecture is presented. Since Google Chrome has a clear trend of growth (Statcounter, 2020) on all types of devices (tablets, mobile phones, laptops, desktops, etc.), an extension for this browser is created. Once the Chrome extension is installed, the user can access a web hyperlink or use a search engine for the topic of interest and click on the extension icon.

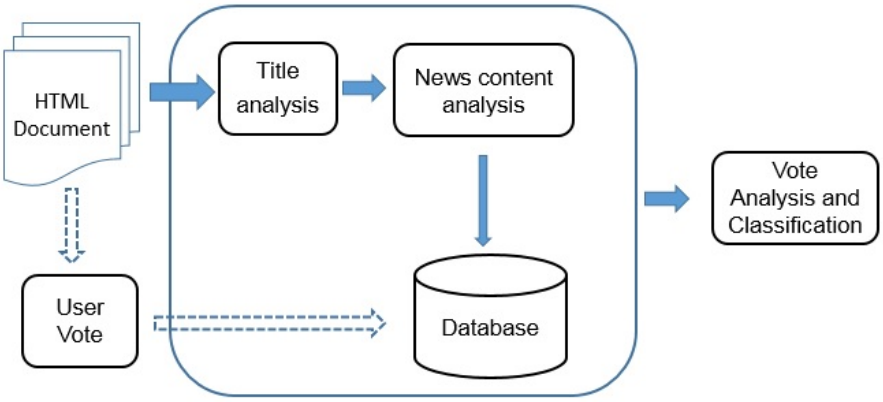
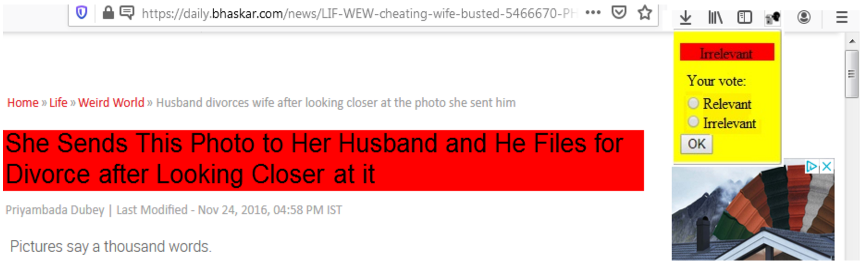


Fig. 2 – The architecture of the proposed framework.

The URL (Uniform Resource Locator) of the accessed page is identified and sent to the server. Using this link, the server tries to identify if the news have been accessed or not. If it already appears in the database, the final answer will be the registered class, considerably reducing the response time. If the URL is not found in the repository, the source page is requested to identify the elements necessary for classification (title, content). After receiving the answer, the background color of the title changes depending on the result obtained. If an irrelevant class is detected the title background is marked in red otherwise it is colored green. In Fig. 3 the classification class for (Bhaskar, 2020) web page is presented.


https://daily.bhaskar.com/news/LIF-WEW-cheating-wife-busted-5466670-PH

Home » Life » Weird World » Husband divorces wife after looking closer at the photo she sent him

She Sends This Photo to Her Husband and He Files for Divorce after Looking Closer at it

Priyambada Dubey | Last Modified - Nov 24, 2016, 04:58 PM IST

Pictures say a thousand words.

Irrelevant

Your vote:

○ Relevant

○ Irrelevant

OK

Fig. 3 – The result of automated analysis and the extension popup.

The title and content of the page will be analyzed and some similarity metrics between the title and the content will be calculated. Obviously, the keywords from the title should be found in the content of the article. Consequently, one of the methods of analysis is counting the occurrences of the keywords in the text of the article, after excluding, of course, the linking words. On the other hand, the user can give a vote after reading the article, thus indicating, from his point of view, if there is a relevant link between the headline and the content. The results of these operations will be stored in a database. The reader's vote is optional, but it can improve the accuracy of the classification. However, this vote is not added to the database unless the user is registered, in order to eliminate possible abuses.

2.1. User Database

Spring is one of the most popular application development framework for Java. A significant number of developers use it to create high-performance codes that are easy to maintain. An important benefit is that it does not need an application server, but creates an embedded one in the application itself. It also facilitates the creation and connection to a H2 database, without the need for an external server (Spring, 2020). H2 is a relational database management system written in Java. It can be embedded in Java applications or run in client-server mode.

In the proposed framework, the H2 database is used to keep the users' data as well as their votes. In order to do this, three table were created (Fig. 4):

USERS table – contains the list of users who have created an account;

NEWS table – contains the title, the links and the descriptions of the news;

EVALUATIONS table – contains the classification given by the users for each news.

The data of this repository can be used to create a model closer to the desire and understanding of the readers. Of course, for the same news, different users may vote in a different way (for someone news may be relevant in relation to its title, for someone else it may not). At the same time, a user can vote only once for an article item, being identified by the IP (Internet Protocol) address of the computer. A jQuery framework function was used to retrieve this address. For security, the server checks that the IP is valid. jQuery is an open-sourced JavaScript library that allows web developers to add extra functionality to their websites.

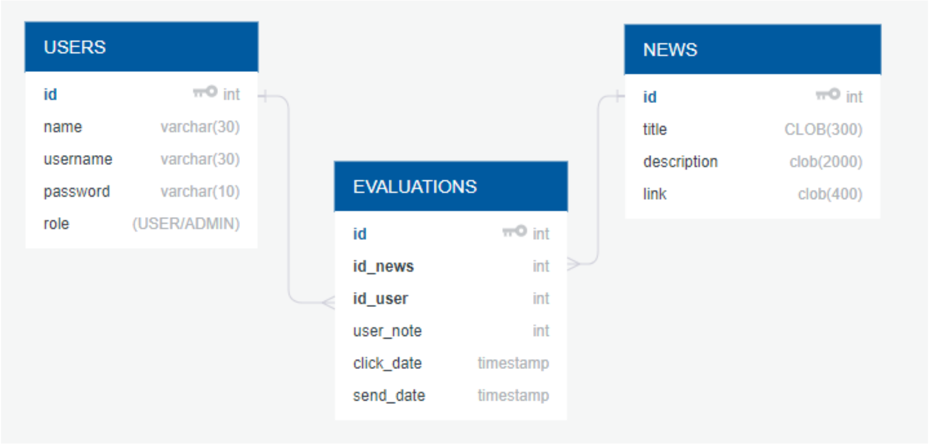


Fig. 4 – The UML (Unified Modeling Language) diagram for used H2 database.

The news can be found with a search engine, but can also be chosen from a list that is updated using RSS (Really Simple Syndication) web feeds. There are a lot of addresses used to download online news, among which: https://www.bzi.ro/; https://www.activenews.ro/; https://www.agerpres.ro; https://www.theguardian.com/us-news/rss; https://news.yahoo.com/rss/us; https://www.bbc.com/news/.

2.2. The Title and the Content Analysis

First, the server withdraws the title from the online article. Typically, a headline is inserted into the web page using the HTML tag <*h1*>, therefore such an element was sought in the page. The text from the first <*h1*> found element is extracted and considered to be the title of the article. However, there are still pages that have several such elements. Thus, it was decided to take the text of the first <*h1*> element. If no <*h1*> tag is found or if the text is empty, the string is taken from the <*title*> element of the web page. Although it is not guaranteed that the extracted string containing the headline is always the correct one, the practice has shown that in most cases it meets the requirements.

The title of the paper usually contains a small number of words, whatever the language is and it should summarize the content of the text as well as possible. However, any title will also contain connecting words, prepositions, interjections, etc., which are not related to the actual analysis of the text. In the literature it is called ***stopwords***. There are several lists of such words on the Internet, the most useful being, as a rule, those used by the Google search engine (CodeGoogle, 2020). A list of stopwords was written in a text file. These words were removing from string extracted from the title. Because this framework was created especially for articles in Romanian, some other operations were performed before text analyzing. The diacritics have been removed because there is no standard to clarify whether the online text is written with or without them, each web page creator writing according to his own desire. Also, all the words were transformed in the analysis stage in lower case, in order to count as accurately as possible the words from the article, including those at the beginning of the sentences or those written only in capital letters.

Starting from the premise that the remaining words in the title are keywords, the appearances of each were counted in the rest of the article. Each news item from the H2 database was analyzed. Since web pages are heterogeneous in structure, even if they are written in HTML, it was necessary to remove as many of those elements that could introduce noise: unnecessary tags (like <header>, <footer>, <aside>, etc.), comments from other users, unnecessary images or links to other pages. The stopwords and the diacritics were also removed from the news content, and the entire text was converted to lowercase.

The Jsoup library was used to download the page and identify the items of interest from its content. Jsoup is a Java library for working with HTML, for fetching URLs and extracting and manipulating data, using the HTML5 DOM (a Document Object Model) methods and CSS selector (Jsoup, 2020).

Following the operations described, the number of occurrences in the text of the relevant words from the title will be inserted in a HashMap. Because it is desired to store this information in a database, MongoDB was used. This facilitates the storage of data structures in BSON (Binary JavaScript Object Notation) format. BSON’s binary structure encodes the type and length of information, which allows it to be parsed much more quickly (Mongodb, 2020).

MongoDB is a document database: each record in a MongoDB collection is a document (Mongodb, 2020). Documents are a structure composed of file and value pairs. Fields in documents can hold rich data: other documents, arrays of values or documents (Mongodb, 2020). In our case, the MongoDB collection contains a document about each news in which records are kept, such as: the title, the hyperlink of the web page, the list of associations between keywords and the number of these words in the article, the number of positive (relevant) and negative (irrelevant) votes calculated from H2 database and the class in which it has already been placed by the previous votes or by the previous classifications.

Fig. 5 describes the steps followed from downloading the web page to counting the keywords in the article.

Step 1
Download title and content from the web page;

Step 2
Transform words into lowercase letters to eliminate cases where the word that appears either at the beginning of a sentence or somewhere in the text, is capitalized;

Step 3
Eliminate diacritics: some web pages use diacritics and others do not; not all pages use UTF-8 (Unicode Transformation Format);

Step 4
Split after any non-alpha-numeric character

Step 5
Create a HashMap <*String, Integer*> in which all the words in the title are stored as a key and corresponding to each word the value 0 is attached. Later, it will be used to store the number of founded keywords in text;

Step 6
Counting the keywords: if a word from the headline is in the text, the corresponding value from the HashMap is increasing;

Step 7
Create a Mongo element: for each association <word, number of occurrences> a mongo type element is created and a list is formed with all of them; this will be useful to store the data in MongoDB.

Step 8
Create a MongoDB.

Fig. 5 – The steps for MongoDB repository development.

2.3. The Feature Extraction and the Classification

Due to the fact that not every user will have enough patience to wait a long time for loading a web page, the issue that the proposed application must answer is to find an efficient classification method that has a high accuracy of results and a response time as short as possible. The use of Mongo technology aims to ensure that the analysis time is as fast as possible. For the same reason, it is desired that the set of training data to be maintained at a relatively small number which can lead to significant classification errors. In H2 database up to 100 news items were stored, but the number can be set by the user with administrator role. If this threshold is exceeded, the oldest news will be removed from the list. Inspired by ensemble methods philosophy, three different ways were used to predict whether or not the headline is relevant to the content of a story. The ensemble methods combine the predictions of several estimators in order to improve the generalization and the robustness over a single estimator (Cîmpanu *et al.*, 2017). Several classifiers were tested: k-Nearest Neighbors (KNN), Random Forest Naïve Bayes. The Random Forest had a very good accuracy rate, but the analysis time was too long (sometimes blocking the browser) while, the Naïve Bayes obtained unsatisfactory error rate results. For this reason, in this work only the data obtained with the KNN algorithm were used.

KNN is a supervised learning algorithm that does not require a training stage. It is based on learning by analogy and determines the class corresponding to a test example based on its similarity to K examples, the most similar, in the training data set (Dumitriu *et al.*, 2018). KNN is preferred when there is no prior knowledge regarding data distribution. Each instance in the training set is a vector in the data representation space and is assigned a single label (class, target, etc.). The training step for the KNN algorithm consists only in storing the feature vectors and the labels corresponding to the classes for the training examples. In this work the user vote labels the training data. The class of a new instance is chosen as the one having the majority vote. Distance measures like Euclidian distance, Minkowski distance, Correlation index, City-Block distance, Cosine similarity function, Hamming distance, and Jaccard index can be used. After several attempts that did not show clear or spectacular differences between the various metrics, the Euclidean distance and Cosine similarity function were preferred for the simplicity of the implementation. The Cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance (for example a certain word appears much more often than in the other document) they could still have a smaller angle between them (smaller angle, higher similarity). Of course, a detailed study might look for subtle differences between all metrics, but that was not the main purpose of this study.

The KNN algorithm compares existing news with a new one and classifies it according to the category of the closest K news. Because the distance is calculated based on the vectors associated with the two MongoDB documents, there is a problem with their length, which is given by the number of keywords. For this reason, the calculation of the distance is done using only the common words between the two vectors. If there is an equal distance between the vector corresponding to the news to be classified and two other documents, but the number of occurrences of the words is different, in order to give a higher priority to those with more common occurrences, the value of the distance was calculated with the relation:

$$d(x,y) = \frac{1}{m}\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

where the x vector represent the current news list of keywords with the number of occurrences in the text, x_i is one value from vector *x*. The vector *y* is one news item from training database, and y_i is one value from vector *y*. The *n* is the number of common keywords found in the both vectors and *m* is the total number of the occurrences of the keywords in both vectors.

Choosing the right value for *K* is a difficult issue. In this work it was considered the largest odd number less than or equal to 10% of the number of records in the training data set. If the size of the list of articles containing common words is less than the value of *K*, it will adjust to the number of news in the list. If there is no list of common words or all values for distances are 0, parts of the words will be searched by removing the last letters of each word. But this process is time consuming.

After this procedure, a first classification (named *VOTE1*) is obtained for the new web page. The second way to classify the same page use Cosine similarity function as a metric:

$$d(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{2}$$

where the *x*, *y*, *n* have the same meaning as in equation (1).

To establish a second classification (*VOTE2*), the weighted sum of all positively classified news (S_x^+) and the weighted sum of negatively classified news (S_x^-) that have common keywords with the new instance were calculated.

$$S_x^+ = \sum_{i=1}^{p}\left(\frac{1}{d(x,y)}w_i\right) \tag{3}$$

$$S_x^- = \sum_{i=1}^{q}\left(\frac{1}{d(x,y)}w_i\right) \tag{4}$$

where *x* represents the current news, *y* represents one news from data training, *p* and *q* are the number of positive respectively negative classified news from the training database which has common keywords with the new instance *x*, *d*(*x*,*y*) is Cosine distance calculated according to Eq. (2), and w_i is the number of common keywords between vectors *x* and *y*. The highest value from S_x^+ and S_x^- indicates the class of the new instance.

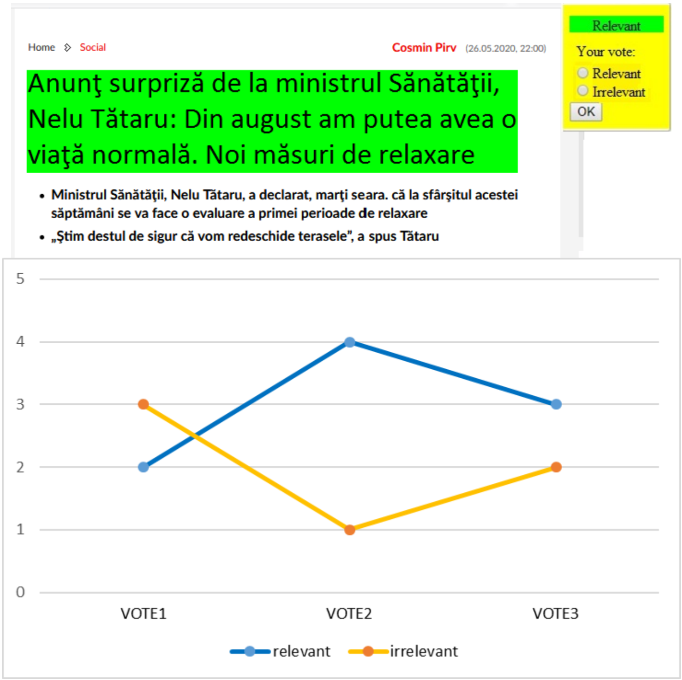
Because these two procedures show sometimes antagonistic classes, a third procedure based on the words occurrence frequency is taken into consideration (*VOTE3*). For each news item having common words with the new instance, the average frequency of keywords appearing in the text is calculated as the sum of all common keywords occurrences divided by the number of keywords. Using this average frequency, the news will be sorted in descending order. In case of a tie, the news with many common words will have priority. The new class is established with the majority vote of first *K* sorted items.

3. Results and Discussions

Some students volunteered to populate the training database. For this, a large number of online news was randomly selected using RSS web feeds from Romanian language news sites among which: https://www.agerpres.ro, https://www.bzi.ro/, https://www.activenews.ro/. Each volunteer was able to select from the news list thus generated, the news that they found attractive and that they chose a class. For some of this news, more than one volunteer voted, for others not. In the end, it was found that 39% of the news considered received positive votes (relevant) and 61% were marked as having an irrelevant title in relation to the content. Later, other adult persons using this application, voted for news that aroused their interest.

A k-fold statistical cross-validation method was used in order to evaluate KNN learning algorithm. To evaluate a predictive model the training dataset is divided between a train dataset and a test set (Refaeilzadeh *et al.*, 2009). During the k-fold cross-validation procedure, the whole set is randomly partitioned into *k* datasets (named folds). Each fold is used one time for the validation of the model obtained when the rest of *k*-1 folds are used as training data. All *k* results from the folds are combined to produce a single estimation. Using k-fold cross-validation method (*k*=10), five runs were performed for each classifier. The average values of the validation errors obtained are: 4.87% for *VOTE1* procedure, 5.15% for *VOTE2* method and 5.36% for *VOTE3*.

In Fig. 6, the vote obtained based on each of the three methods can be seen. Because two of them are positive, the news items were classified as positive, so that the title is considered relevant in relation to the content.



Home ⮚ Social Cosmin Pirv (26.05.2020, 22:00)

Anunţ surpriză de la ministrul Sănătăţii, Nelu Tătaru: Din august am putea avea o viaţă normală. Noi măsuri de relaxare

- Ministrul Sănătăţii, Nelu Tătaru, a declarat, marţi seara. că la sfârşitul acestei săptămâni se va face o evaluare a primei perioade de relaxare
- „Ştim destul de sigur că vom redeschide terasele", a spus Tătaru

Relevant

Your vote:
○ Relevant
○ Irrelevant
OK

Fig. 6 – A web page and the corresponding votes obtained based on the three methods.

When other users accessed the application, there was an average of 15% false positive news reported (Fig. 7), and 9% false negative in relation to the subsequent vote given by the users (Fig. 8).

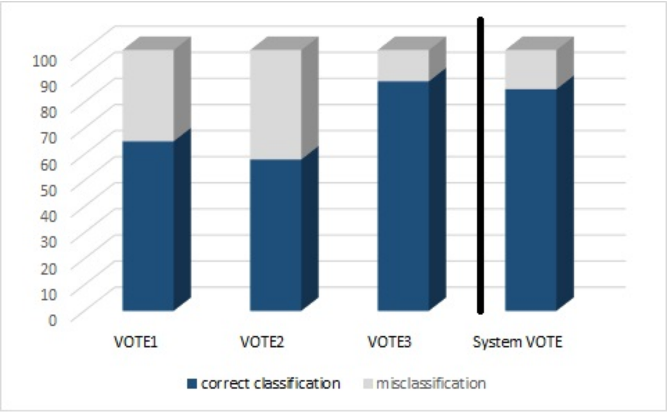


Fig. 7 – The average value of the vote percentages in case of positive classification.

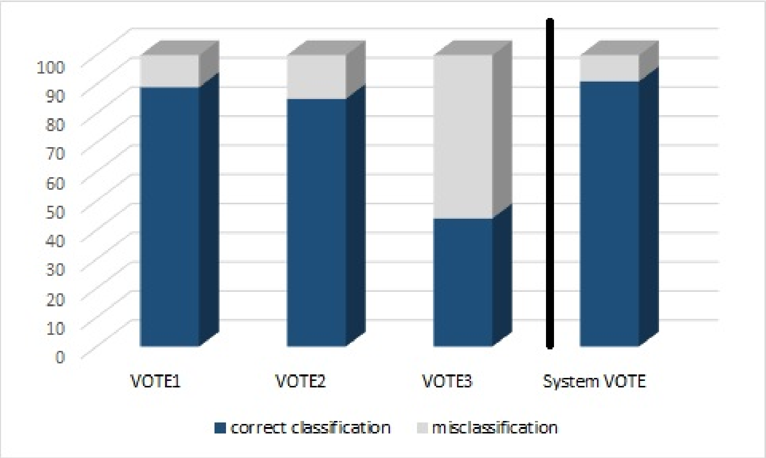


Fig. 8 – The average value of the vote percentages for the irrelevant class.

4. Conclusions

The purpose of the present study is the design and implementation of a framework for detection of misleading headlines, which is an increasingly pressing issue especially in the online environment. Thus, a method of analyzing the titles and their content is proposed so that the detection accuracy of the articles meant to attract only through their title to be as good as possible.

A KNN classifier and some well-known metrics were used in order to achieve this goal. A dynamic and fast method of analysis using MongoDB database was studied. An extension for Google Chrome browser has been created in order to facilitate the use of the application, but this method can be easily adapted to other browsers. The user has the opportunity to access any web address and to activate the application according to his desire. The framework can assist him in deciding whether or not to read a particular news item.

The study of other classification methods, an efficient method for choosing the best value of K or the use of other metrics for KNN should empower this framework. Also, the words analysis for removing suffixes and prefixes and obtaining the roots of words to find more precisely and correctly similarities should increase the efficiency of the proposed method.

The tests carried out proved that the proposed method can be used successfully in detecting misleading titles.

Acknowledgements. The authors wish to thank the volunteer students and all collaborators for their involvement in this study.

REFERENCES

Agrawal A., *Clickbait Detection Using Deep Learning*, 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), Dehradun, 268-272 (2016).

Allcott H., Gentzkow M., *Social Media and Fake News in the 2016 Election*, Journal of Economic Perspectives, **31**, *2*, 211-236 (2017).

Cîmpanu C., Ferariu L., Dumitriu T., Ungureanu F., *Multi-Objective Optimization of Feature Selection Procedure for EEG Signals Classification*, Proc. of 6th Edition of the International Conference on e-Health and Bioengineering, EHB 2017, 22-24 June, Sinaia, Romania, 434-437 (2017).

Dumitriu T., Cimpanu C., Ungureanu F., Manta V.-I., *Experimental Analysis of Emotion Classification Techniques*, Proceedings of 2018 IEEE 14th International Conference on Intelligent Computer Communication and Processing, ICCP 2018, 63-70 (2018).

Figueira Á., Oliveira L., *The Current State of Fake News: Challenges and Opportunities*, Procedia Computer Sience, **121**, 817-825 (2017).

Kaur S., Kumar P., Kumaraguru P., *Detecting Clickbaits Using Two-Phase Hybrid CNN-LSTM Biterm Model*, Expert Systems with Applications, 151, (2020).

Matraguna M., *Clickbait: You won't Believe what it Really is* (transl. from Roumanian), https://tutoriale-pe.net/ce-este-clickbait/ (last updated 2017), last visit May 2020.

Molyneux L., Coddington M., *Aggregation, Clickbait and Their Effect on Perceptions of Journalistic Credibility and Quality*, Journalism Practice, **14**, *4*, 429-446 (2020).

Munger K., *All the News That's Fit to Click: The Economics of Clickbait Media*, Political Communication, **37**, *3*, 376-397 (2020).

Pengnate S., *Shocking Secret you won't Believe! Emotional Arousal in Clickbait Headlines: An Eye-Tracking Analysis*, Online Information Review, **43**, *7*, 1136-1150 (2019).

Refaeilzadeh P., Tang L., Liu H., *Cross-Validation*, In Encyclopedia of Database Systems, Springer, 259-265 (2009).

Scacco J.M., Muddiman A., *The Curiosity Effect: Information Seeking in the Contemporary News Environment*, New Media and Society, **22**, *3*, 429-448 (2020).

Xu P., Wu C.-S., Madotto A., Fung P., *Clickbait? Sensational Headline Generation with Auto-Tuned Reinforcement Learning*, Proceedings of the Conference The 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, 3065-3075 (2020).

*** Cambridge Online Dictionary, https://dictionary.cambridge.org, last visit May 2020.

*** Chrome Web Store, https://chrome.google.com/webstore, last visit May 2020.

*** https://code.google.com/archive/p/stop-words/, last visit May 2020.

*** https://daily.bhaskar.com/news/LIF-WEW-cheating-wife-busted-5466670-PHO.html, last visit May 2020.

*** https://jsoup.org, last visit May 2020.

*** https://www.mongodb.com, last visit May 2020.

*** https://www.naijaloaded.com.ng/entertainment/you-wont-believe-it-50-year-old-mother-of-3-breaks-the-internet-with-birthday-photoshoot-photo, last visit May 2020.

*** https://gs.statcounter.com, last visit May 2020.

*** Spring Framework – Overview, https://www.tutorialspoint.com/spring/, last visit May 2020.

ANALIZA ŞTIRILOR ONLINE PENTRU DETECŢIA TITLURILOR ÎNŞELĂTOARE - STUDIU DE CAZ

(Rezumat)

Mass-media contemporană creează un mediu favorabil pentru distribuirea de ştiri false. Identificarea automată a acestui tip de ştiri este subiectul multor dezbateri şi încercări de a dezvolta programe de calculator. Un caz particular al acestui fenomen este reprezentat de titlurile atractive, care induc cititorului o idee greşită despre conţinutul articolului. Această lucrare propune o aplicaţie prin care eventualele ştiri online, cu o mare discrepanţă între titlu şi conţinut, pot fi identificate şi semnalate cititorului. Algoritmul k-Nearest Neighbors şi unele metrici pentru măsurarea similarităţilor au fost studiate cu scopul de a identifica dacă titlul este relevant în raport cu conţinutul articolului. Limbajul Java a fost folosit pentru procesarea textelor şi pentru a analiza frecvenţa cuvintelor din titlu şi aspectarea acestora în conţinutul ştirii. Rezultatele sugerează că aceste metode pot fi utile pentru detectarea titlurilor înşelătoare.