Application of TextRank algorithm for credibility assessment

Bartłomiej Balcerzak[1], Wojciech Jaworski[1,2], and Adam Wierzbicki[1]

[1]Polish-Japanese Institute of Information Technology, Koszykowa 86, 02-008 Warsaw, Poland,
[2]Institute of Informatics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland, wjaworski@mimuw.edu.pl,

***Abstract*—In this article we examine the use of TextRank algorithm for identifying web content credibility. TextRank has come to be a widely applied method for automated text summarization. In our research we apply it to see how well does it fare in recognizing credible statements from a given corpus. So far, research into use of NLP algorithms in credibility assessment was focused more on extracting the most informative statements, or dealing with recognizing the relation between claims within a document. In our paper, we use a collection of 100 websites reviewed by human subjects in regard to their credibility, therefore allowing us to check the algorithm's performance in this task. The data collected showed that the TextRank algorithm can be used for recognizing credibility on the level of aggregated statement credibility.**

Keywords: Text Summarization, TextRank, Information Credibility, Web 2.0

I. Introduction

In the ever growing plethora of content available on-line, identifying credible claims and telling them apart from frauds becomes nearly impossible. Discerning the truthfulness of statements, a task difficult in itself, becomes even harder when the amount of data that needs to be reviewed is as enormous as the World Wide Web. It is reasonable to argue that text summarization tools such as TextRank may serve as a method of automating the task and providing filters for credible and informative data. In this paper we are going to test the TextRank algorithm as a method for identifying both credible statements. By using data collected from human respondents we can check how well does the algorithm fare compared to manually assigned ranks. However the most important point of our research is to investigate the limitations of using this method for credibility assessment. In the following sections we will first briefly describe the algorithm itself. Afterwards we are going to present our hypothesis regarding the possibilities of using TextRank for identifying credible content. In the later section we will analyze the empirical set used for testing the said hypothesis, and discuss the results put forward. The last part will represent conclusions derived from our work.

II. Related work

Although the TextRank is a widely implemented algorithm used for text summarization, it was not widely used in research concerning web content credibility. Most of its application involved keyword and sentence extraction. Research into this was mostly done by [4] [5]. It showed that TextRank is an efficient method for text extraction, that does not require human supervision, nor does it need to be language specific [6]. It was shown also to outperform many supervised methods. Using this tool for credibility assessment is a fresh field of investigation, so far only few studies involving this have been produced. One notable approach to text summarization with TextRank and content credibility was attempted by [3]. In their study, the authors use TextRank algorithm as a supplement to STATEMENT MAP tool [8], which was designed to extract and resolve conflict between statements within one document. TextRank was used as one of the algorithms that would select statements further review. In another study conducted by [2] a more direct link between web content credibility and text summarization algorithms was established. By using data from Tweeter authors wanted to verify the hypothesis connecting content informativeness and credibility that was originally proposed by [1]. They managed to confirm the hypothesis for heterogeneous graphs of users and tweets. To the best of our knowledge no attempts at direct empirical testing of TextRank's performance in identifying credible sentences.

III. TextRank algorithm

In this section, we give a brief description of the TextRank algorithm and it's attributes. TextRank is an unsupervised algorithm, used for automated summarization of texts written in natural languages. It belongs to the category of extractive summarization techniques, that is, it extracts the most important sentences from the original text and uses them to construct a summary. It can be used both for extracting keywords from a corpus of words or sentences from a body of documents. It uses a graph-based approach in which sentences/words are considered vertices of a graph, and the similarity between them is equivalent to the weighted edges of the said graph. This is an expansion of the PageRank algorithm [7] and it utilizes its modified formula used for calculating scores for each sentence or word. The formal expression of the algorithm itself and the similarity function it uses can be found in [5], [6] The TextRank algorithm has some major advantages which make it a natural choice for using in extractive summarization:

- **It is unsupervised**, therefore no set of training data, moreover no human generated input (i.e. a manually tagged corpus) is needed for it to process actual data.
- **It is language independent**. The TextRank algorithm is based only on word concurrence and does not require

any knowledge of the grammar. This excludes the need for creating tools dedicated to particular languages. However the algorithm may be limited by different ways of separating sentences in various languages, for example due to different alphabets, or standards of writing.
- **It is well defined and developed.** At the moment many implementations of the algorithm exist, making it easily available for developers, who wish to utilize its properties.

In this paper we use a TextRank algorithm implementation based on an indirected graph. The weight applied to edges of the said graph is equivalent to cosine similarity measure. Assigned scores will be used as prediction of the sentences' credibility.

If TextRank would prove useful in credibility assessment it could provide a great tool for web content credibility enhancement. In this paper we use the basic TextRank settings to check how well they fare in the task of identifying credible sentences.

In our analysis we will use two implementations of the TextRank algorithm for comparison with the ranks collected from humans:

- **100 separate corpora.** In this approach the algorithm is run for each of the documents separately, therefore each individual website serves as it's own corpus. This method is used primarily to test the algorithm's capability to recognize the most informative and credible statements.
- **1 combined corpus.** For this application we merged all of the documents into one large corpus, on which the algorithm was run. This method is used primarily to obtain a combined ranking of all sentences, allowing us to compute the average importance and credibility score of particular sentences.

Both approaches are also used for testing whether the size of the corpus affects the algorithm's performance.

IV. Hypothesis and measures used

In this section we present the main hypothesis that we aim to verify in the course of this paper:

Statements selected with the use of TextRank algorithm are more likely to be credible than those statements that have been randomly selected

By testing this hypothesis we want to check whether there exists a connection between centrality calculated by the TextRank algorithm and the perception of credibility. The base method of testing this hypothesis would be to pick n-top sentences or documents, from those selected by TextRank and calculate the frequency of n-top credible statements or documents. This parameter would be then compared with a similar measure derived from a comparison of n statements and documents selected at random. If that hypothesis held true this would mean that the basic assumption of TextRank allows for applying the algorithm in credibility assessment. The following variables have been utilized to verify this hypothesis:

- Credibility ratings given by human subjects
- Importance ratings given by TextRank - 100 corpora implementation
- Importance ratings given by TextRank - 1 merged corpus implementation

The following measure has been used for testing the hypothesis:

- **Precision.** This is a classic measure in studies of text summarization techniques, although its origins trace from information retrieval. In the case of our research it is the percentage of documents and sentences that attained in both top n ranks from both human ranks and those given by the algorithm. This measure is most important if we are interested in quality of identifying the most viable information. For the sake of analysis three distinct cut off values were chosen: top 30%, top 25% and top 20% (used only for whole documents). When analysing sentences, precision was calculated for each document separately, this would be later aggregated and presented as general value of Precision.

We also wanted to see the relation between human generated importance and credibility.

V. Experimental Data

In our research we decided to use a collection of 100 texts from web pages grouped in four categories. In the course of our papers those 100 texts will be referred to as documents. The four groups in which the documents were divided are:

- Personal Finance (58 web pages)
- Healthy lifestyle (28 web pages)
- Politics (9 web pages)
- Entertainment (5 web pages)

Each page was divided into separate sentences. Eight evaluators would rate the informative importance and credibility of each sentence on a 0 to 100 scale (where 0 - not informative/credible, 100 - highly informative/credible). The total number of sentences was 1592, which gives the average number of sentences per document equal to 15,92.

Based on those ratings an aggregated score (arithmetical mean) for each sentence was calculated. Afterwards we calculated a mean credibility score for a sentence in each document. This served as a basis for human generated ranks used in further analysis.

Afterwards both TextRank implementations were programmed with the use of Python language. With all the stopwords and punctuation points filtered out, the algorithm has been run. Scores assigned by the program were used to rank the sentences, and average scores were calculated for whole documents, similarly to what was done for human generated ranks.

VI. Results

As shown in table I, human subjects highly correlated sentence importance with its credibility. Precision exceeds 60% suggesting a strong connection between the two aspects

TABLE I
HUMAN GENERATED IMPORTANCE VS. CREDIBILITY - SENTENCES

Precision (top 30 %)	Precision (top 25 %)
61%	52%

TABLE II
HUMAN GENERATED IMPORTANCE VS. CREDIBILITY - DOCUMENTS

Precision (top 30%)	Precision (top 25%)	Precision (top 20%)
40 %	40%	43%

of a sentence in human evaluation. Data gathered from whole documents presents a connection weaker than that evident with sentence rating. Still these finding corroborate what was already shown.

Table III shows that TextRank achieved modest result in recognizing the most important sentences in a document. As it can be seen from comparison of table III and table IV the difference in performance between the 100 corpora and 1 merged corpus method are minimal if not negligible. It may seem that the size of the corpus did not have a significant effect on the algorithm.

As it was the case with finding informative statements, TextRank produced a relatively modest outcome. Obtained values for precision (see table V and VI) are significantly lower than those attained by humans in the experimental situation. This may lead to a conclusion that while basic TextRank algorithm works better than random selection there is still room for improvement. In spite of that, it has to be pointed out that human scores exhibited a strong bias towards connecting credibility with informativeness. This means that the basic TextRank algorithm, as the one used in our research, may be used, to some extent, as a method of extracting an unbiased selection of sentences for further manual review.

Analysis of TextRank performance with identifying documents paints a slightly different picture. The scores are higher and are more or less consistent for all the cut-off points. Each of the scores for importance is higher that a Precision taken from a ranking system generated at random (30,2%, 25%, 22% respectively). What is most interesting however are the scores achieved when we look at mean credibility value for a sentence in a document. The values produced are only slightly lower than those produced by human subjects. It also showed the most significant results for the lowest cut-off value of top 20 (35% as compared to 43% for human ranks and 22% for ranks generated at random). This may indicate that TextRank algorithm may be successfully used for extracting the most credible documents, based on their informativeness. This further reinforces the hypothesis set forth at the beginning of our paper.

TABLE III
HUMAN GENERATED IMPORTANCE VS. TEXTRANK 100 CORPORA

Precision (top 30 %)	Precision (top 25 %)
41 %	32%

TABLE IV
HUMAN GENERATED IMPORTANCE VS. TEXTRANK 1 CORPUS

Precision (top 30 %)	Precision (top 25 %)
42 %	32%

TABLE V
HUMAN GENERATED CREDIBILITY VS. TEXTRANK 100 CORPORA

Precision (top 30 %)	Precision (top 25 %)
37 %	29%

TABLE VI
HUMAN GENERATED CREDIBILITY VS. TEXTRANK 1 CORPUS

Precision (top 30 %)	Precision (top 25 %)
35 %	27,5%

TABLE VII
HUMAN GENERATED IMPORTANCE VS TEXTRANK - DOCUMENTS

Precision (top 30 %)	Precision (top 25 %)	Precision (top 20 %)
47%	43%	40%

TABLE VIII
HUMAN GENERATED CREDIBILITY VS TEXTRANK - DOCUMENTS

Precision (top 30 %)	Precision (top 25 %)	Precision (top 20 %)
34%	32%	35%

VII. CONCLUSIONS AND OBSERVATIONS

Data collected and analysed in our research leads as to the following conclusions regarding the use TextRank for credibility assessment:

- In general, TextRank algorithm achieved better results than a random process of selection. This confirms the hypothesis put forward at the beginning of our paper. Although TextRank produced modest results when it came to emulate the performance of human subjects at identifying the most credible statements, it holds great promise as a tool for aiding manual credibility assessment, since it provides an extractive summary which not only include same of the most informative statements, but also selects those that can be omitted due to human bias, and since experimental data has shown that people tend to correlate importance with credibility, the algorithm may serve as a method of overcoming this bias when preparing data for manual assessment. When used for identifying credible and informative documents, TextRank fared relatively better in selecting the most informative documents. It was also comparable to human evaluator when using importance rankings as a method of predicting mean sentence credibility per a given document.

  There are also some observations we made:

- Human subject tended to strongly correlate importance of particular statements with their credibility. The sentences deemed the most informative for a single document where also mostly ranked as relatively credible.
- When considering TextRank performance in identifying the most informative statements within documents, no larger difference has been observed between the 100 corpora and 1 merged corpus approach. Even though the latter method showed as improvement of the former the gain is not significant enough.

To sum up, our research has produced data suggesting, that for human evaluators the concepts of credibility and importance are highly connected, leading to a suggestion that a larger bias may be at work. In this aspect, our man hypothesis has been validated. The algorithm was not successful at extracting the most credible statements. In this context our sub-hypothesis can not be considered true. Regardless, the same program was more efficient when dealing with documents as a whole. Still, more work is needed in this field of investigation.

VIII. Future work

The study shown that more research is needed in the field of using automated text summarization techniques in credibility assessment. In our future work we plan to expand it by using larger corpora dedicated to single topics, and more refined weighting functions such as *uniqueness* or *fungability* as proposed by [3]. Extracting the most unique sentences may show to be a more effective method of extracting passages for credibility assessment. We also plan to employ different unsupervised and not language specific tools in order to see if the could outperform the basic TextRank algorithm. Data collected so far lead us to suggest that the idea of centrality, a hub of TextRank, can be useful in credibility assessment. Therefore we would like to implement solution that still utilizes this approach. Moreover in this study we did not investigate the influence of controversy an the algorithms performance. In some topics, such as politics or fringe science this factor me significantly distort the results given by human participants. An algorithm taking this problem into account could produce better results. Such method was to some extent used by [3]. However they were mostly focused on resolving conflicts between varying statements within a document. Using Text Summarization tools when dealing with matters of controversy opens a new field of inquiry into human perception of controversy itself, and the persistence of cognitive biases when reviewing web content. In summary, the field of automated credibility assessment holds great promise, in our future work we plan on further developing the tools needed to improve the process.

IX. Acknowledgments

This work was financially supported by the European Community from the European Social Fund within the *INTERKADRA* project.

It was also supported by the grant *Reconcile: Robust Online Credibility Evaluation of Web Content* from Switzerland through the Swiss Contribution to the enlarged European Union.

References

[1] Duan, Yajuan, et al. ”An empirical study on learning to rank of tweets.” Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, 2010.

[2] Huang, Hongzhao, et al. ”Tweet Ranking Based on Heterogeneous Networks.” COLING. 2012.

[3] Kaneko, Koichi, et al. ”Mediatory Summary Generation: Summary-Passage Extraction for Information Credibility on the Web.” PACLIC. 2009.

[4] Mihalcea, Rada, and Paul Tarau. ”TextRank: Bringing order into texts.” Proceedings of EMNLP. Vol. 4. No. 4. 2004.

[5] Mihalcea, Rada. ”Graph-based ranking algorithms for sentence extraction, applied to text summarization.” Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Association for Computational Linguistics, 2004.

[6] Mihalcea, Rada. ”Language independent extractive summarization.” Proceedings of the ACL 2005 on Interactive poster and demonstration sessions. Association for Computational Linguistics, 2005.

[7] Page, Lawrence, et al. ”The PageRank citation ranking: Bringing order to the web.” (1999).

[8] Murakami, Koji, et al. ”Statement map: reducing web information credibility noise through opinion classification.” Proceedings of the fourth workshop on Analytics for noisy unstructured text data. ACM, 2010.