Implementing Fact-Checking in Journalistic Articles Shared on Social Media in the Philippines Using Knowledge Graphs

Donata D. Acula[1], Louise Aster C. Oblan[2], Tracy B. Pedroso[3], Katrine Jee V. Riosa[4], Michelle Arianne R. Tolibas[5]
Institute of Information and Computing Sciences
University of Santo Tomas
Manila, Philippines
e-mail: {ddacula[1]; louiseaster.oblan.iics[2]; tracy.pedroso.iics[3]; katrinejee.riosa.iics[4]; michellearianne.tolibas.iics[5]}@ust.edu.ph

***Abstract*—In the technology age, articles with fraudulent content are rampant, especially articles shared on social media. Misinformation could just be an inaccuracy at its best, or it could lead to normalizing false information at worst. To aid the predicament, the researchers created a system that will "fact check" suspicious articles against those articles that have been deemed credible, reliable, and more accurate, in order to help fight deceiving content that may be detrimental to society. The journal regarding computational fact checking that was published by Ciampaglia, et. al. (2015) from the Indiana University in the USA entitled Computational Fact Checking from Knowledge Networks, was used as the basis and inspiration for this thesis. The researchers made use of the undirected graph (UG) together with a part-of-speech (POS) tagging algorithm to create a knowledge graph (KG) that would serve as the center of the system. Five different POS tagging algorithms were paired with the UG to assess which combination would yield the best results, these are Conditional Random Fields, Logistic Regression, a Hybrid of CRF and LR, Random Forests, and K-Nearest Neighbors. Random Forests and K-Nearest Neighbors were classification algorithms used in Ciampaglia's study. It was concluded that among the 5 pairs of UG and POS Tagging algorithms, the Hybrid of CRF and LR used as a POS tagger, together with the UG, created the most efficient KG.**

Keywords-knowledge graph; undirected graph; conditional random fields; logistic regression; hybrid; random forests; k-nearest neighbors; part-of-speech tagger; fact checking

I. INTRODUCTION

Fact-checking is very crucial not just in online news industry. Reporting for everyday citizens that seek new questions and demand fact-checked information can be challenging especially when correcting recently posted false articles. It needs to be more persuasive and believable than the false article posted [1]. Fortunately, GL Ciampaglia's team from Indiana University in the US recently created a computational algorithm intended for fact checking [2].

They claim that the important and complex human task of fact checking can be effectively reduced to a simple network analysis problem, which is easy to solve computationally. With the integration of knowledge graphs, the team from IU was able to achieve exceptional results and thus, making it the first process that was able to have computational fact-checking study [2]. With the algorithms and established processes from said study as the primary source and basis of this thesis, the team proposed to integrate Pattern Recognition and Natural Language Processing in order to bring veracity and integrity to the information on the internet.

II. BACKGROUND OF THE STUDY

For many years, journalists struggle to find a true and reliable source. Sometimes, some of their information has unwanted false facts that may destroy their credibility as writers and may also influence or supply their readers with the wrong information they presented. Due to the reason that journalists write as soon as they detect the information, there is only minimal time to fact-check and cross-reference their information with other journalists, thus, making them vulnerable to misinformation and inaccuracy.

These kinds of false journalism are popular on many social networking sites such as Facebook, Twitter and/or Tumblr and people sometimes are blindly following or believing on a false article that is beautifully written [3].

With these problems at hand, the team developed easier and more optimal algorithms that can fact-check and cross-reference articles and posts that may appear on the timeline/newsfeed of the user.

III. OBJECTIVES

The main objective of the study was to improve the system of Ciampaglia's team by combining Natural Language Processing and Pattern Recognition algorithms that will reduce the information lost when using a knowledge graph that is undirected, and evaluated using simple shortest path when fact checking. It specifically aimed to:

1. create a better system than the system of Ciampaglia's team using the combination of Natural Language Processing and Pattern Recognition algorithms to create the undirected Knowledge Graphs;
2. prove the superiority of the new system in terms of running time and space; and
3. compare the overall efficiency of the new system to the method of Ciampaglia, et al.

978-1-5386-6350-9/18/$31.00 ©2018 IEEE

IV. PROPOSED AND CURRENT SOLUTION

A. Input

The input of the training set is the specific site (www.philstar.com) in which all the articles are deemed true. The input of the testing set, on the other hand, is the article to be fact-checked and the input of the training set in order to compare both the inputs and determine the accuracy of the article to be checked.

B. Processes

In order for both systems to accept the validity of the article, two criteria should be met: validity of the POS-tagging algorithms and the validity of the undirected graph.

The system starts by crawling the news website (www.philstar.com) and retrieving each link pertaining to different categories (see Figure 1). These articles contained in each section were saved in the database, under its corresponding category.

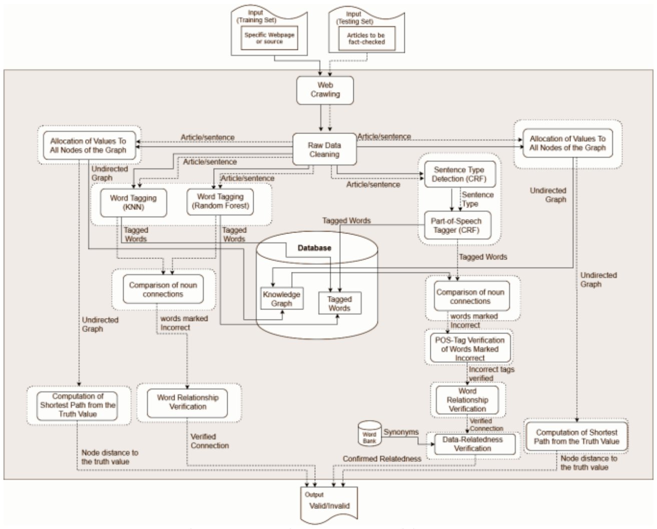
Input (Training Set)
Specific Webpage or source
Input (Testing Set)
Articles to be fact-checked
Web Crawling
Raw Data Cleaning
Article/sentence
Article/sentence
Allocation of Values To All Nodes of the Graph
Allocation of Values To All Nodes of the Graph
Undirected Graph
Article/sentence
Article/sentence
Article/sentence
Word Tagging (KNN)
Word Tagging (Random Forest)
Sentence Type Detection (CRF)
Sentence Type
Part-of-Speech Tagger (CRF)
Undirected Graph
Tagged Words
Tagged Words
Tagged Words
Tagged Words
Database
Knowledge Graph
Tagged Words
Comparison of noun connections
Comparison of noun connections
Undirected Graph
Undirected Graph
words marked Incorrect
words marked Incorrect
POS-Tag Verification of Words Marked Incorrect
Incorrect tags verified
Word Relationship Verification
Word Relationship Verification
Verified Connection
Computation of Shortest Path from the Truth Value
Computation of Shortest Path from the Truth Value
Word Bank
Synonyms
Data-Relatedness Verification
Verified Connection
Confirmed Relatedness
Node distance to the truth value
Node distance to the truth value
Output
Valid/Invalid

Figure 1. Main system architecture.

To get more accurate results, cleaning the data from unnecessary characters was done and then the data was transformed into an input of five modules for the Part-Of-Speech tagging: the K-Nearest Neighbor and Random Forest algorithms of the old system (left), the Conditional Random Field (CRF) of the new system (right), and the creation of the Knowledge Graph in both new and old systems.

The cleaned data for both the training and the testing dataset goes to the new system's CRF which identifies the sentence type in order to help the part-of-speech tagger to identify each word correctly from each article. This was done in order to determine the most appropriate part of speech that a word belongs to. However, in its counterpart on the old system, both the KNN and the RF only tagged the words and has no sentence type detection.

After CRF tagged each word in the articles, the training data goes directly into the database where it is stored (see Figure 2). The testing data, on the other hand, goes to the Boyer-Moore Algorithm which extracts the tagged nouns and matches it in the data of the training set in order to compare the information that was connected into them (see Figure 3).

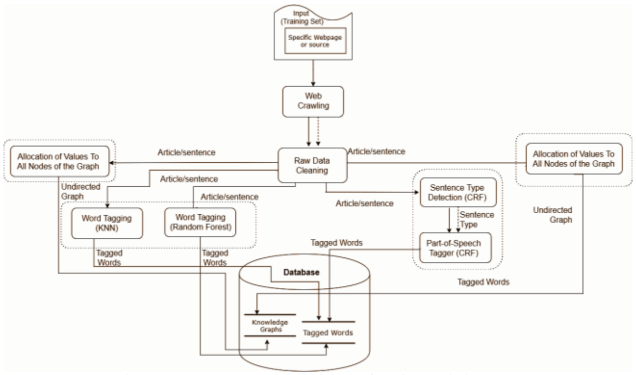

Figure 2. System architecture for the training set.

Since the assumption in the data from the training set is always true, the algorithm, then, determines if the information in the testing data is true or not based on the information found on the training data. The same thing was done in the old system, with KNN and RF, separately instead of CRF.

After the Boyer-Moore searches the best fit article for the test set data, the erroneous words that was not found by Boyer-Moore in the new system was re-verified by Logistic Regression (LogReg) and was tagged again; it is because the later findings show that the accuracy for LogReg exceeds that of CRF's accuracy. With that, the new system compares again the relationship of the words in both training and testing using Semantic Relatedness (SR), same with that of the old system's Semantic Proximity.

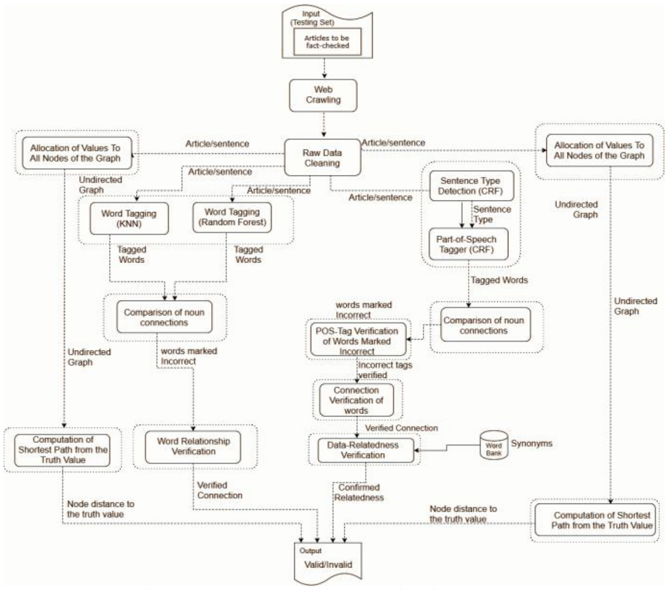
Input (Testing Set)
Articles to be fact-checked
Web Crawling
Allocation of Values To All Nodes of the Graph
Article/sentence
Raw Data Cleaning
Article/sentence
Allocation of Values To All Nodes of the Graph
Undirected Graph
Article/sentence
Article/sentence
Article/sentence
Sentence Type Detection (CRF)
Sentence Type
Word Tagging (KNN)
Word Tagging (Random Forest)
Undirected Graph
Part-of-Speech Tagger (CRF)
Tagged Words
Tagged Words
Tagged Words
Undirected Graph
words marked Incorrect
Comparison of noun connections
POS-Tag Verification of Words Marked Incorrect
Comparison of noun connections
Undirected Graph
words marked Incorrect
Incorrect tags verified
Connection Verification of words
Verified Connection
Computation of Shortest Path from the Truth Value
Word Relationship Verification
Data-Relatedness Verification
Word Bank
Synonyms
Verified Connection
Confirmed Relatedness
Node distance to the truth value
Node distance to the truth value
Computation of Shortest Path from the Truth Value
Output
Valid/Invalid

Figure 3. System Architecture for the Test Set

However, the words of different articles with same content are not exactly the same. The authors of the articles may differ in their wordings and sentence instruction and with this, there is a possibility that an article from the testing data may be deemed as invalid even if it conveys the same information to that of an article in the training data.

In an attempt to rule out this problem, if the SR of the new system did not find the relationship in the words, the Harmony Search Algorithm (HSA) searches for possible synonymy of the information in both the training and testing data. The HSA, with the help of a word bank, finds all possible synonyms in the every word that is incorrectly tagged and assesses them with the words found in the training data, it searches and optimizes the best fit article for the input. After the HSA confirms or rejects the synonymy, only then can the system partially decide if the testing article is valid or not.

Another criterion on both systems, as mentioned before, is the creation of knowledge graph (KG) and finding its shortest path. After accepting the data from the cleaning, the creation of both systems' knowledge graph which allocates the words into nodes and connects them based on their relatedness in the article.

C. Output

In order to determine the validity of the articles to be fact-checked, the new system has two criteria:

1. The validation of the shortest path of the testing data to the truth value in the training data; and
2. The confirmed relatedness of the testing data to the training data, using the Harmony Search Algorithm (HSA) to compare the synonymy of the sentences in the training and testing data.

With these, the researchers established that if the results of the two criteria are met, the article checked is valid. However, if one (i.e., The HSA confirmed the relatedness but the shortest path is greater than the sum of the word count and 25% of the said count (count + (0.25*count); or the shortest path is less than the said sum but the HSA did not confirm the relatedness) or both of the criteria failed, the article checked is invalid. On the other hand, an article will only have a 0% validity if no related article was found in the database. It does not mean that the article tested is 0% true. With this, the researchers included another output which displays the percentage validity of an article, or the percentage that matches in the system's database (which is presumed to be true/factual).

V. Statistical Treatments and Test Results

A. Normality of the Data

The number of test data that were crawled by the system is more than 40,000. To correctly test the significant difference of the systems, the normality of the dataset was first tested with:

$$Skewness = \frac{3(\bar{x}-M_o)}{s} \qquad (1)$$

For the dataset to be normal, the skewness of the data should be in 1 to -1 in range [4]. With this, the researchers tested the normality of the data using two categories: the year they got published and their section classification (e.g., business, politics, etc.) and the results for both categories came as positively skewed (1.218 for the year and 1.101 for the classification). Because of this, the researchers used the non-parametric tests to prove the significant difference between the systems.

B. Differences between Algorithms

In order for the system to have an accurate output for the related articles and most related article, the precision of the data were calculated. By doing this, the values for true positive, true negative, false negative and true negative were tested using 150 articles.

TABLE I. Tallies of Conditions Negative and Positive for Precision and Accuracy

	RF	KNN	CRF	LogReg	Hybrid
True Positive	49	99	101	99	***102***
False Negative	53	3	1	44	***0***
True Negative	***48***	47	44	3	44
False Positive	***0***	1	4	4	4

Basing on the tallies (see Table I), this indicates that if the algorithms of the new system are to compare with that of the algorithms in the old system, this shows that there is a difference between the old system and the new system in terms of Memory Usage and Running Time.

TABLE II. Overall Precision and Accuracy of the Algorithms

Algorithms	CRF	LogReg	Hybrid	KNN	RF
Accuracy	0.519	0.618	***0.913***	0.514	0.325
Precision	0.537	0.632	***0.849***	0.512	0.240

The overall precision and accuracy of the algorithms were calculated based on the tallies indicated in Table I (see Table II).

C. Wilcoxon Signed Rank Test

The Wilcoxon Signed Rank Test was used to determine the significant difference between the old system and the new in terms of memory usage and running time. The difference of the old (current) and new (proposed) systems were with W⁻ where:

$$W^{-} = \Sigma(ranks\ of\ negative\ differences) \qquad (2)$$

The two systems were compared in terms of their POS-tagging accuracy, running time and space occupied (see Table III).

TABLE III. Z-Values of the New System Versus Old System in Terms of Running Time and Memory with Alpha = 0.05

Algorithms	KNN		Random Forest	
	Running Time	Memory Usage	Running Time	Memory Usage
CRF	-6.962	-2.19	-9.699	-1.724
Logistic Regression	-10.505	-0.538	-10.505	-0.661
Hybrid	-10.668	-8.389	-11.901	-9.247

The shortcut to hypothesis testing of the Wilcoxon signed rank-test is the critical z-value for a 95% confidence interval which is z=1.96 for two tailed and directionality [5].

With this, the researchers computed the means of the differences of the new system and old in terms of their memory usage and memory. Table III shows that the new system is significantly better in terms of memory usage and memory than the old system since the z-value of each algorithm used in the new system are all less than 1.64.

D. McNemar's Test

For the overall accuracy, McNemar's Test was used. It is a non-parametric measure for paired nominal data. McNemar test was also used to find if there was a change in proportion of the data. It is a modification of the ordinary chi-square test that takes the paired nature of the responses into account [6]. The results of the old and new system to the 120 true articles that were tested were used. The discordants were used to calculate the test statistic. If the p-value is significant or less than the alpha of 0.05, the null hypothesis will be rejected.

TABLE IV. P-Values of the New System Versus Old System In Terms of Accuracy with Alpha = 0.05

Algorithms	KNN	Random Forest
CRF	<0.000001	<0.000001
Logistic Regression	0.363	<0.000001
Hybrid	0.0195	<0.000001

The researchers used only articles that came from the database since both the systems should accept it as valid. Each algorithm was then assessed based on their outputs on whether they validated the articles or not. Table IV shows that the new system is significantly better in terms of running time than the old system since all of the z-value for each algorithm used in the new system is less than the 1.64.

E. Cochran's Q Test

As for the precision, an extension of McNemar's Test which is Cochran's Q Test was used. Multiple pairwise comparisons were applied using the McNemar procedure for the precision [7].

TABLE V. P-Values of the New System Versus Old System In Terms Of Accuracy With Alpha = 0.05

Algorithms	KNN	Random Forest
CRF	<0.0001	<0.0001
Logistic Regression	<0.0001	<0.0001
Hybrid	<0.0001	<0.0001

Using the Cochran's Q test, the algorithms' precision used in the old and new system was compared. Since all the computed p-value in Table V is less than α and is almost 0, then it is statistically proven that there is a significant difference between the new and old system.

VI. Conclusions, and Recommendations

A. Conclusions

The researchers concluded that the new system is significantly better compared to the Ciampaglia's. According to the tests made, there is a significant difference in terms of accuracy of the algorithms of the new system compared to the old system, with the Hybrid of Logistic Regression and CRF has the highest accuracy followed by Logistic Regression and CRF.

In terms of running time and memory, it can be seen that the Hybrid of Logistic Regression and CRF is also significantly better compared to the other algorithms, although CRF performed worse compared to Random Forest and KNN.

Overall, the researchers concluded that the new fact-checking system is significantly better than the system of Ciampaglia with the tests conducted, with the combination of Logistic Regression and CRF as the best among the other tested algorithm in this study, in terms of accuracy, runtime and space.

B. Recommendations

The researchers recommend the following for future studies regarding this paper:

1. Try using different algorithms to pair with the undirected graph to possibly improve the new system's performance.
2. Operate the system with the help of an external server to possibly improve running time.
3. Add more reliable sources in the training data to have wider scope in fact-checking.
4. Find out a system which should recognize foreign words in order to possibly improve the accuracy of each algorithm.

Acknowledgment

First of all, we are grateful to the Lord, for giving us the blessings of patience, intelligence and most of all, clarity of mind that made us possible to conduct every action to complete this paper. We thank Him for guiding us every step of the way. To our panel members, Mr. Khristamonte M. Balmeo, Mr. Joseph Richard G. Catubag and Assoc. Prof. Perla P. Cosme, and our coordinator Asst. Prof. Cecil Jose A. Delfinado, we give our deepest gratitude to them for finding time to give us their observations and expertise with regard to our study. With these, our study has improved a lot. To our linguists, Dr. Redentor S. Rodriguez, Mr. Jay Mark A. Balason, and Ms. Bianca Chia R. Castro, we want to thank them for finding the time to work with us in all their busy and cramped schedules, in order to make our system better and more reliable. We also express our deepest gratitude to our family and friends especially to Mr. Brian Daniel L. Tiongco, Ms. Cyrinne Jazmine D. Abenir, and Ms. Alyanna M. Delos Santos, for supporting, helping and believing in us whenever we feel down. We thank them for always being there for us in every way that they could, for always cheering

us up. We hope we made them proud by making this study. We love you all!

References

[1] Aldabe, I., Bogaar T., Erp, M. V., Fokkens, A., Ploeger, T., Rigau, G., Rospochera, M., Soroa, A., Vossen, P. (2016) Building Event-Centric Knowledge Graphs from News. Journal of Web Semantics Vol (1). pp. 1-5

[2] Bollen, J., Ciampaglia, G. L., Flammini, A., Menczer, F., Rocha, L., Shiralkal, P. (2015). Computational Fact-Checking from Knowledge Networks. Public Library of Science, Indiana University, USA.

[3] Alejandro, J., (2010). Social Media: Web 2.0 and the News. Journalism in the Age of Social Media. pp. 10

[4] Rouse, M. (2017). Skewness. Retrieved from: http://whatis.techtarget.com/definition/skewness

[5] Glen, S. (2015). Wilcoxon Signed Rank Test: Definition, How to Run. Retrieved from: http://www.statisticshowto.com/wilcoxon-signed-rank-test/

[6] Petrie, A., Watson, P.F. (2010). Method Agreement Analysis: A Review of Correct Methodology. Elsevier Inc. Theriogenology.

[7] Glen, S. (2016). What is Cochran's Q Test? Retrieved from: http://www.statisticshowto.com/cochrans-q-test/