Research Article

Fakeheader: A Tool to Detect Deceptive Online News Based on Misleading News Headlines and Contents

Normala, Che Eembi @ Jamil[1], Iskandar, Ishak[2], Fatimah Sidi[3], Lilly Suriani, Affendey[4]

[1,2,3,4]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400, UPM Serdang, Selangor Malaysia
iskandar_i@upm.edu.my[2]

Article History: Received: 10 November 2020; Revised: 12 January 2021; Accepted: 27 January 2021; Published online: 05 April 2021

Abstract: Online news has been the primary source of news content for newsreaders. Unfortunately, based on several findings, readers tend to judge on specific events based on the news headlines rather than its contents. With the advancement of mobile and web technologies, it is easier to spread the news to others with these unhealthy habits that can cause negative impacts on individuals, organizations, or nations that are victimized by the news. In the proposed work, a tool to detect deceptive news based on misleading headlines or content is developed. The tools incorporate data veracity framework for online news with Support Vector Machine and proposed combination of features. The experimental results show the proposed tool managed to produce high performance results with more than 90% precisions and recalls.

Keywords: deceptive news, misleading headlines, deception detection, detection tool

1. Introduction

The Internet has been the source for the users to search almost everything for their daily life. They searched the Internet for the things they want buy, travel, socialize, banking and financing as well as finding reading materials (Malviya, 2010) (Iskandar Ishak et al 2012) (Sidi et al 2013) (Saad et al 2014; 2016) (Alwan et al 2016; 2017)). News is among the most accessible reading materials that people will read and discuss daily, but researchers have found out that trustworthiness, and the truth of the news content can be lacking (Osgood, 1971); (Knapp, Hart, & Dennis, 1974); (Dirsehan & Çelik, 2011); (Al-Kinani et al. 2020); (Jamil et al. 2015); (Kerby & Marland, 2015). Similarly, online news shares the same problem, but the scale of the problem is even greater than the traditional paper-based news as online news can be shared rapidly through computers and mobile devices regardless of its trustworthiness.

Some literatures also highlighted the impacts of deceptive writing. As an example, it is found that when information are hidden with added cognitive for deception, it changes in human behaviour form (Frank et al., 2008). Jung reported about false details through the media can influence receivers (Jung, 2009). In news reporting, the headline is one of the critical parts of the news report authors. It provides the fundamental idea of the news, and it allows readers to choose from a large number of news items in which they summarized the content of the story through the title. However, there were times that 'catchy headline' approach is used to get the reader's attention while the content is totally or partially different that its headlines. Plus, the media always manipulated the use of the title as an attention grabber to increase their news rating (Dor, 2003) (Ecker, U.K, Lewandowsky, S., Chang, E.P., Pillai, 2014)(D. Q. Wang, 2016). Therefore, it is very important to have a tool to detect news that are misleading and fake in which it is also the objective of this research.

2. Materials and Methods

Some initial success in deception detection approaches has created a new wave of applying intelligent technologies to support deception detection on fake news (Zhou, Shi, Zhang, & Sears, 2006); (Lukoianova & Rubin, 2013); (V. Rubin, Conroy, Rubin, et al., 2016); (Ali et al. 2020); (Ruchansky et al., 2017); (Shu et al., 2017); (W. Y. Wang, 2017); (Karimi & Tang, 2019); (K.-C. Yang et al., 2019); (Yoon et al., 2018); (Esteves et al., 2019). However, most of the previous approaches did not focus on misleading headlines, did not acknowledge the news data structure that contains header, content and other metadata or having low detection accuracy. Since headlines are deemed as critical part of the news, deception detection approach must acknowledge the news data structure in order to detect misleading headlines.

From the previous studies, researchers have utilized numbers of features combination. These features are essential to train the data to determine deception detection classifiers. There are a few features that have been highlighted in the previous researches. Among the features utilized were Absurdity and Humour, Punctuation, Grammar, Body-independent feature, Body-dependent feature, N-gram, Cosine Similarity, and Deception

Detection measurement. In the proposed approach, a new set of combination of features is proposed. In this approach, Bigram and Lemmatization features are proposed to be combined with the Base (TFIDF), Syntactic, Bigrams (N-Grams) and Punctuation features to produce prediction technique with high accuracy (precision, recall and F-score). Figure 1 shows the proposed framework for detecting deceptive news based on misleading headlines or contents using the proposed combination of features. In this study, the dataset from Fake News Dataset (FN) (McIntire, G., 2018) has been selected to validate the proposed approach. The dataset contains 6,335 articles. 3,171 of them are labelled as real news and 3,164 of them are labelled as fake news. The ratio of real and fake news articles in the dataset are around 1:1 in which the titles, contents and veracity labels are provided. The dataset is grouped into three types; news headlines, news content (without headline), and combined (headline+content).

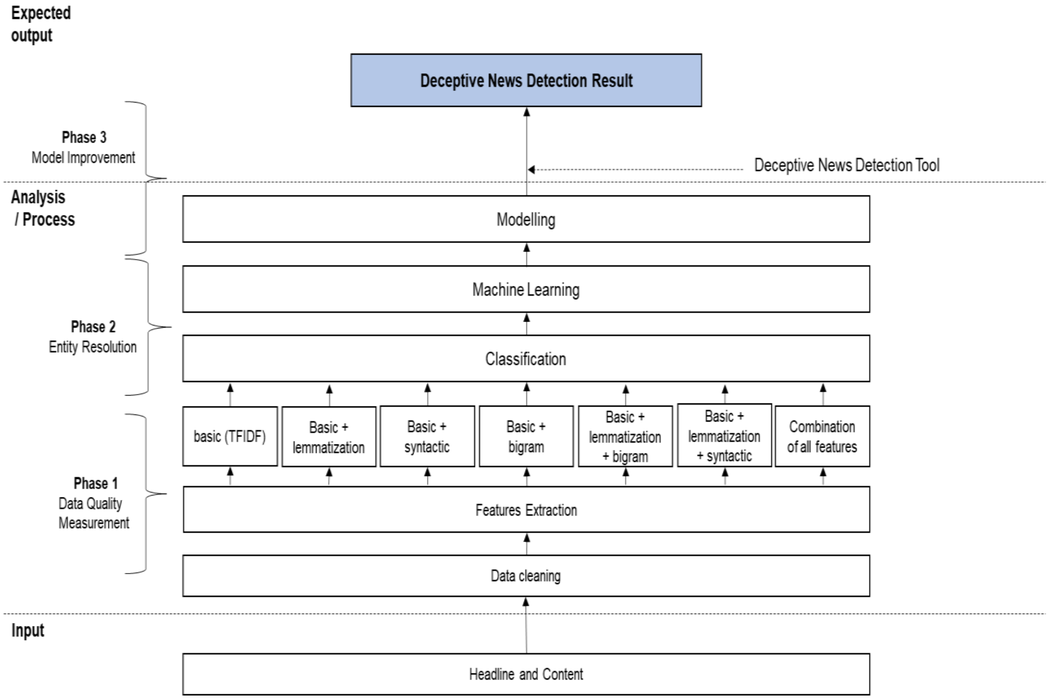


Figure 1. Data Veracity Framework for Detecting Deceptive News based on Misleading Headlines

3. Results and Discussion

A number of experiments are conducted using 80% of the dataset for training and 20% for testing with five-fold cross-validation. Subsequently, the classification technique tested with five different types of base classifiers, namely Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), Logistic Regression (LR), K-Nearest Neighbour (kNN) and Artifical Neural Network (ANN) applied at the training stage. The parameter is set up under Linear Kernel SVM with probability TRUE and C=5. SGD used loss parameter 'hinge,' penalty l2, alpha, and tol are $le - 3$, and maximum iteration 1000. LR used parameter solver 'lbgs' and maximum iteration of 1000.

Based on the training results shown in Table 1, 2 and 3, SVM classifier emerged as the best classifier with highest accuracy in majority numbers of features used with more than 90% accuracy recorded for all types of datasets used. SVM topped as the classifier with the highest accuracy in detecting fake news over the headline dataset for all features used. SVM also topped as the classifier with highest accuracy over the content dataset for all features except for Base + Syntactic feature where ANN record the highest accuracy. For combined dataset (headline + content), SVM classifier recorded highest accuracy on all features except for Base+Lemmatization, Base+Syntactic, and All features. ANN classifier recorded the highest accuracy for Base+Lemmatization, Base+Syntactic, Base+Lemma+Syntactic and All features.

Table 1. Accuracies on headlines dataset

Features	***SVM***	***SGD***	***LR***	***k-NN***	***ANN***
Base (TFIDF)	**87.45%**	77.86%	81.08%	75.60%	81.34%
Base + Lemmatization	**93.77%**	90.84%	92.38%	71.82%	93.41%
Base + Syntactic	**82.09%**	72.29%	81.61%	77.71%	82.31%

Base + Bigram	**93.51%**	91.68%	92.93%	88.03%	92.74%
Base+ Lemma + Syntactic	**94.18%**	90.89%	93.31%	82.86%	94.06%
Base + Lemma + Bigrams	**93.51%**	91.71%	92.93%	88.03%	92.74%
All	**93.94%**	92.07%	92.88%	84.32%	92.98%

Table 2. Accuracies on content dataset

Features	*SVM*	*SGD*	*LR*	*k-NN*	ANN
Base (TFIDF)	**94.81%**	74.38%	78.82%	77.45%	77.96%
Base + Lemmatization	**97.95%**	94.81%	96.59%	55.86%	97.81%
Base + Syntactic	79.83%	65.07%	78.87%	78.34%	**83.63%**
Base + Bigram	**98.53%**	93.44%	96.64%	89.02%	48.75%
Base + Lemma + Syntactic	**97.55%**	91.39%	94.86%	81.42%	98.05%
Base + Lemma + Bigrams	**98.53%**	94.06%	96.63%	89.06%	48.75%
All	98.34%	87.26%	92.76%	85.79%	**98.73%**

Table 3. Accuracies on combined dataset

Features	*SVM*	*SGD*	*LR*	*k-NN*	ANN
Base (TFIDF)	**94.29%**	73.45%	78.73%	77.61%	85.58%
Base + Lemmatization	98.17%	95.50%	96.76%	86.36%	**97.81%**
Base + Syntactic	79.49%	62.85%	78.68%	77.20%	**83.77%**
Base + Bigram	**99.35%**	94.46%	96.54%	87.99%	99.16%
Base + Lemma + Syntactic	**97.64%**	91.06%	94.71%	81.15%	**97.64%**
Base + Lemma + Bigrams	**98.70%**	94.54%	96.54%	87.99%	99.16%
All	98.52%	87.27%	92.90%	85.06%	**98.73%**

Based on the training experiments, SVM topped the accuracy on all types of dataset and SVM classifier is chosen to be included into the implementation of deception detection tool of online news based on misleading headlines.

3.1 Performance Measures

In order to evaluate our approach, experiments have been implemented to detect misleading online news by using the Fake News Dataset. Table 4,5 and 6 present the measures of precision, recall, and F-score with associated five-fold cross-validation results for our deception detection model. Each table represents different data type. Based on the results, the proposed approach recorded high precision for Base+Bigram, Base+Lemma+Bigram and Base+Lemma+Syntactic features (with 98% and 99% precision). This shows that the proposed approach with the above features produced high precision of prediction on headline dataset, which generally are short texts. In terms of Content dataset (without headline), similar features recorded among the highest precision as well as Recalls and F-Scores. Content dataset has longer text; therefore, recall and F-score were also high. This shows that the proposed approach with proposed combination of features able to predict deceptive news with high accuracy. In terms of combined dataset (Headline+Content), similar features recorded high accuracies, recalls and F-Scores (except for Base+Lemma+Syntactic feature). In general, the proposed approach with combination of features that mainly consist of Bigram and Lemmatization recorded high precision. As the dataset grows in size (Content and Combined dataset), the approach recorded high precision as well as recalls and F-Scores.

Table 4. Performance of dataset (Headline)

Features	*Precision*	*Recall*	*F-Score*
Base (TFIDF)	89	85	87
Base + Lemmatization	90	90	94
Base + Bigram	**99**	89	94
Base + Syntactic	79	83	81
Base +Lemma+ Bigram	**99**	89	94
Base + Lemma + Syntactic	**98**	**98**	**98**
All	97	91	94

Table 5. Performance of dataset (Content)

Features	*Precision*	*Recall*	*F-Score*
Base (TFIDF)	91	99	94
Base + Lemmatization	**98**	**98**	**98**
Base + Bigram	**98**	**98**	**98**
Base + Syntactic	75	82	78
Base +Lemma+ Bigram	**98**	**99**	**98**
Base + Lemma + Syntactic	**98**	97	97
All	**98**	**98**	**98**

Table 6. Performance of dataset (Headline + Content)

Features	*Precision*	*Recall*	*F-Score*
Base (TFIDF)	91	98	94
Base + Lemmatization	**98**	**98**	**98**
Base + Bigram	**98**	**99**	**99**
Base + Syntactic	75	82	78
Base +Lemma+ Bigram	**98**	**99**	**99**
Base + Lemma + Syntactic	91	90	91
All	**98**	**99**	**99**

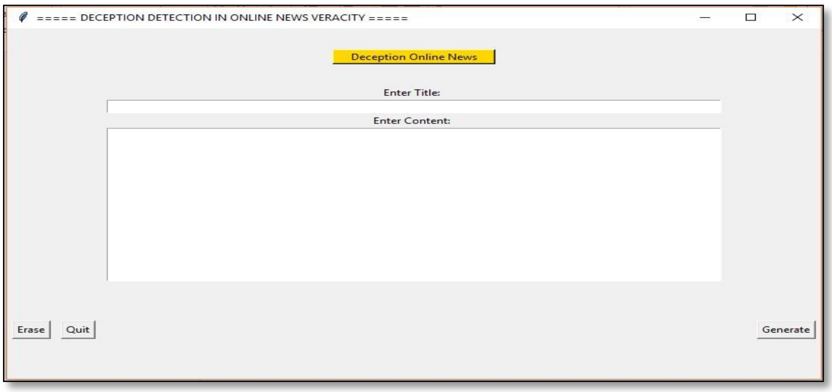
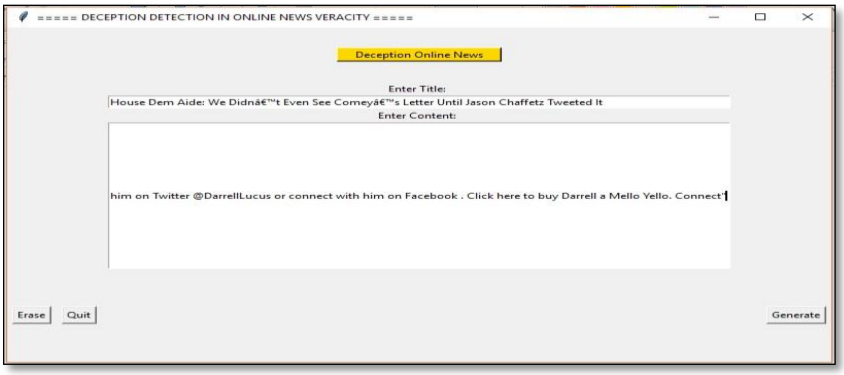


Figure 2. FakeHeader Input Interface

3.2 FakeHeader development

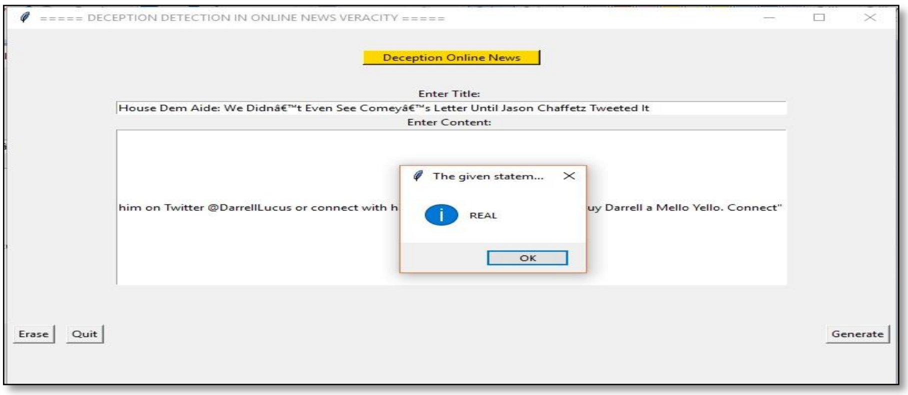
FakeHeader is developed as a prototype tool that can be used to detect fake news using news headlines, contents or the combination of both. The tool is developed using Python version 3.7.3 on Windows 10 Professional by using 64-bit as an operating system. Below is the Graphical User Interface (GUI) created for the proposed tool. In order to detect deceptive news, users can input either title of the news, content of the news or the whole news into the text area to get the prediction and accuracy of the fake news detection result. Figure 2, 3, 4 and 5 show the developed interfaces of FakeHeader.



House Dem Aide: We Didnâ€™t Even See Comeyâ€™s Letter Until Jason Chaffetz Tweeted It

him on Twitter @DarrellLucus or connect with him on Facebook . Click here to buy Darrell a Mello Yello. Connect'

Figure 3. FakeHeader Data Input Interface with input text



===== DECEPTION DETECTION IN ONLINE NEWS VERACITY =====

Deception Online News

Enter Title:

House Dem Aide: We Didnâ€™t Even See Comeyâ€™s Letter Until Jason Chaffetz Tweeted It

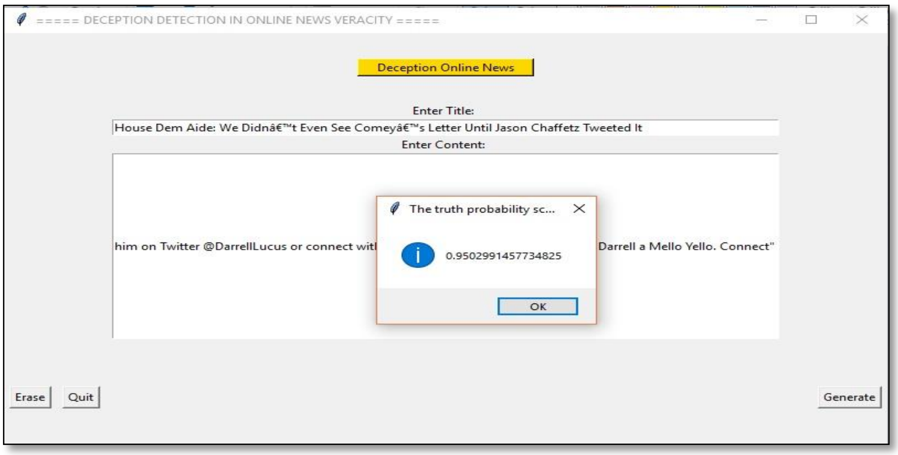
Enter Content:

him on Twitter @DarrellLucus or connect with h ... uy Darrell a Mello Yello. Connect"

The given statem...

REAL

Figure 4. FakeHeader generated results



===== DECEPTION DETECTION IN ONLINE NEWS VERACITY =====

Deception Online News

Enter Title:

House Dem Aide: We Didnâ€™t Even See Comeyâ€™s Letter Until Jason Chaffetz Tweeted It

Enter Content:

him on Twitter @DarrellLucus or connect wit ... Darrell a Mello Yello. Connect"

The truth probability sc...

0.9502991457734825

Figure 5. FakeHeader Accuracy

4. Conclusion

As a conclusion, this paper proposes a simple-to-use tool called FakeHeader to detect deceptive online news based on misleading news headlines, contents or both. The proposed tool incorporates a specialized data veracity framework for data veracity of online news based on news headlines and uses Support Vector Machines classifier. Based on the experiments, the proposed tool with the proposed combination of features scored high accuracy for detecting deceptive news based on headlines, contents and combined news data. For future works, the proposed tool will be validated further with further experiments using datasets that are larger in size for larger scale detection and to improve accuracy.

5. Acknowledgment

Authors acknowledge Universiti Putra Malaysia and Faculty of Computer Science and Information Technology, Malaysia, for all the support. This research is funded by Ministry of Education of Malaysia under the Fundamental Research Grant Scheme (FRGS)(FRGS/1/2015/ICT04/UPM/02/07) and Universiti Putra Malaysia under the Putra Grant Scheme-Putra Graduate Initiative (GP-IPS/2016/9478900).

References

1. Ali, A.S., Zaaba, Z.F., Singh, M.M., Hussain, A. (2020). Readability of websites security privacy policies: A survey on text content and readers. International Journal of Advanced Science and Technology, 29 (6 Special Issue), pp. 1661-1672.
2. Alwan, A. A., Ibrahim, H., Udzir, N. I., & Sidi, F. (2016). An efficient approach for processing skyline queries in incomplete multidimensional database. Arabian Journal for Science and Engineering, 41(8), 2927-2943. doi:10.1007/s13369-016-2048-z.

3. Alwan, A. A., Ibrahim, H., Udzir, N. I., & Sidi, F. (2017). Processing skyline queries in incomplete distributed databases. Journal of Intelligent Information Systems, 48(2), 399-420. doi:10.1007/s10844-016-0419-2.
4. Al-Kinani, M.N.H., Adetunmbi, S.B., Hussain, A. (2020). Usability testing of mobile flipboard application on both non-users and novice users. International Journal of Interactive Mobile Technologies, 14 (5), pp. 47-56.
5. Dirsehan, T., & Çelik, M. (2011). Profiling online consumers according to their experiences with a special focus on social dimension. Procedia - Social and Behavioral Sciences, 24, 401–412. https://doi.org/10.1016/j.sbspro.2011.09.040
6. Dor, D. (2003). On newspaper headlines as relevance optimizers. Journal of Pragmatics, 35(5), 695–721. https://doi.org/10.1016/S0378-2166(02)00134-0
7. Ecker, U.K, Lewandowsky, S., Chang, E.P., Pillai, R. (2014). The Effects of Subtle Misinformation in News Headlines. Uma Ética Para Quantos?, XXXIII(2), 81–87. https://doi.org/10.1007/s13398-014-0173-7.2
8. Frank, M. G., Menasco, M. A., & O'Sullivan, M. (2008). Human behavior and deception detection. Wiley Handbook of Science and Technology for Homeland Security. https://doi.org/0470087927
9. IskandarIshak, Sidi Fatimah, Jabar Marzanah, Sani, Nor Fazlida Mohd, Mustapha, Aida, Supian, S.R. & Apau, M.N.. (2012). A survey on security awareness among social networking users in Malaysia. Australian Journal of Basic and Applied Sciences. 6. 23-29.
10. Jamil, N. B. C. E., Ishak, I. B., Sidi, F., Affendey, L. S., & Mamat, A. (2015). A systematic review on the profiling of digital news portal for big data veracity. Procedia Computer Science, 72 390-397. doi:10.1016/j.procs.2015.12.154.
11. Jung, H. M. (2009). Information Manipulation Through the Media. Journal of Media Economics, 22(4), 188–210. https://doi.org/10.1080/08997760903375886.
12. Karimi, H., & Tang, J. (2019). *Learning Hierarchical Discourse-level Structure for Fake News Detection*. Retrieved from http://arxiv.org/abs/1903.07389
13. Kerby, M., & Marland, A. (2015). *Media Management in a Small Polity : Political Elites' Synchronized Calls to Regional Talk Radio and Attempted Manipulation of Public Opinion Polls*. (August). https://doi.org/10.1080/10584609.2014.947449
14. Knapp, M., Hart, R., & Dennis, H. (1974). An exploration of deception as a communication construct. *Human Communication …*, *Fall*(1), 15–29. https://doi.org/10.1111/j.1468-2958.1974.tb00250.x
15. Lukoianova, T., & Rubin, V. L. (2013). Veracity roadmap: Is big data objective, truthful and credible? *Advances in Classification Research Online*, *24*, 4–15. https://doi.org/10.7152/acro.v24i1.14671
16. Malviya, V. (2010). The Impact of Internet and Digital Media on Reading Habit.
17. McIntire, G. (2018). Fake real news dataset. In George McIntire's Github.
18. Osgood, C. E. (1971). Where Do Sentences Come From? *Semantics, and Interdisciplinary Reader in Philosophy, Linguistics, and Psychology*, 88–105.
19. Rubin, V., Conroy, N. J., Rubin, V. L., Conroy, N. J., Chen, Y., & Cornwell, S. (2016). *Fake News or Truth ? Using Satirical Cues to Detect Potentially Misleading News Fake News or Truth ? Using Satirical Cues to Detect Potentially Misleading News* . (April). https://doi.org/10.18653/v1/W16-0802
20. Ruchansky, N., Seo, S., & Liu, Y. (2017). *CSI: A Hybrid Deep Model for Fake News Detection*. https://doi.org/10.1145/3132847.3132877
21. Saad, N. H. M., Ibrahim, H., Alwan, A. A., Sidi, F., & Yaakob, R. (2014). A framework for evaluating skyline query over uncertain autonomous databases. Procedia Computer Science, 29 1546-1556. doi:10.1016/j.procs.2014.05.140.
22. Saad, N. H. M., Ibrahim, H., Sidi, F., Yaakob, R., & Alwan, A. A. (2016). Computing range skyline query on uncertain dimension, doi:10.1007/978-3-319-44406-2_31.
23. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). *Fake News Detection on Social Media: A Data Mining Perspective*. (i). https://doi.org/10.1145/3137597.3137600
24. Sidi, F., Jabar, M.A., Mustapha, A., Sani, N.F., Ishak, I., & Supian, S.R. (2013). Measuring computer security awareness on internet banking and shopping for internet users. Journal of Theoretical and Applied Information Technology, 53(2): 210-216.
25. Wang, D. Q. (2016). *Madness in the Media : Understanding How People With Lived Experience Interpret Newspaper Headlines*. (April). Retrieved from http://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=5265&context=etd%5Cn

26. Yang, K.-C., Niven, T., & Kao, H.-Y. (2019). Fake News Detection as Natural Language Inference. *Wsdm 2019*. https://doi.org/10.1145/1122445.1122456
27. Yoon, S., Park, K., Shin, J., Lim, H., Won, S., Cha, M., & Jung, K. (2018). *Detecting Incongruity Between News Headline and Body Text via a Deep Hierarchical Encoder*. Retrieved from http://arxiv.org/abs/1811.07066
28. Zhou, L., Shi, Y., Zhang, D., & Sears, A. (2006). Discovering Cues to Error Detection in Speech Recognition Output: A User-Centered Approach. *Journal of Management Information Systems*, *22*(4), 237–270. https://doi.org/10.2753/MIS0742-1222220409