International Journal of Engineering & Technology, 7 (4.31) (2018) 242-245

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Study on Detecting Misleading Online News Using Bigram and Cosine Similarity

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

Faculty of Computer Science and Information Technology
University Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia
**Corresponding author E-mail:normala.jamil@yahoo.com*

Abstract

Fake news can impact negatively in terms of creating negative perception towards business, organization, and government. One of the ways that fake news is created is through deceptive news writing. Many researchers have developed approaches in detecting deceptive news content using machine-learning approach and each of the approach has its own focus. Previous researches emphasis on the components of the news content such as in–detecting grammar, humor, punctuation, body-dependent and body-independent features. In this paper, a new approach in detecting deceptive news based on misleading news has been developed which is focusing on the similarity between the content and its headlines using bigram and cosine similarity. Based on the experiments, the proposed approach has better performance in terms of detecting deceptive news.

Keywords: *Fake news, Deception, Lies, Misleading headlines, Deceiving news*

1. Introduction

In Big Data domain, data veracity is a way to find the truthfulness, availability, accountability and authenticity [1][2][3]. It is an important characteristic that can determine the truth about information being spread to the public such as information from online news. One of the veracity or in our case truthfulness issue in online information such as online news is the spread of news that contains untruth information or fake news. Online news can be categories as unstructured text data. It contains link website, date, day, times, name of author, type of news, video, image and etc. These are text-based information to get to know their customer better and can be used to predict data in future for business decisions and text analytics.

Fake news is defined as news article that are intentionally and verifiably false that could misled readers [4][5][6][7][8][9]. Its widespread can impact negatively in many aspects such as in elections [9][10][11], business [12] and war [9]. Deception is intentional action by an adversary to influence the perceptions, decisions, or actions of the recipient to the advantage of the deceiver [13].

Misleading headline is an example of fake news or deceptive news approach. The trend of users reading only the headlines and the existence of news with misleading headlines are mentioned in many literatures [14][15][16]. Hence, deception detection for online news is very important. Based on the literatures, there are very few approaches being developed to tackle this issue and improving the accuracy of deception detection for online news is the aim of this paper.

Based on the literatures [8][17][5][18], natural language processing (NLP) methods is incorporated with machine learning techniques to identify news that contain deceptive element through its content directly by detecting language pattern, sentiment and word occurrence which is common to news online.

However, based on readers' trend [16][19][14], readers tend to just read the headlines of the news and assumed the whole content of the news based on its headline [14]. In recent years, very few of deception detection approach focuses on the determination of deceptive news of fake news using headlines. Therefore, there is a need a combine approach for comprehensive deception detection for online news by incorporating relevant NLP and machine learning technique.

In order to describe veracity three dimensions are used which are objectivity/subjectivity, truthfulness/deception and credibility/implausibility [18]. The dimensions can reduce noise and potentials error in textual big data due to minimization of bias, intentional misinformation and implausibility. In this paper, we want to highlight the accuracy of the deception in online news based on data veracity.

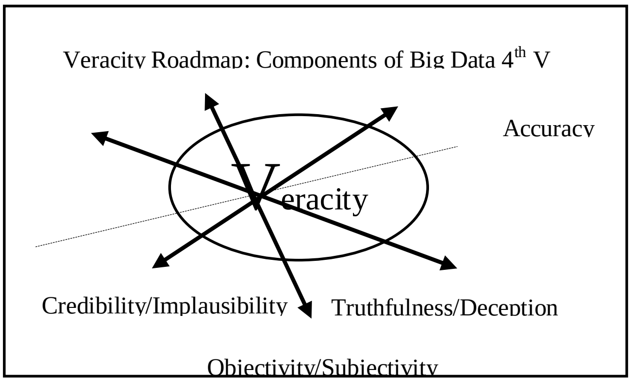

Veracity Roadmap: Components of Big Data 4th V
Accuracy
Veracity
Credibility/Implausibility
Truthfulness/Deception
Objectivity/Subjectivity

Fig. 1: Veracity Roadmap: Component of Big Data 4th V

Previous research papers highlighted issues in detecting deception in online news. It will discuss in the literature review

2. Literature Review

The issues related to identifying truth and to differentiate it from deceptive online news has gained increased attention of researchers especially in data and information management domain [18]. News headlines are an important part in online news reporting and it however can be used negatively to report fake news. According to Journalism and Communication [20], it can classified into two categories of headlines which are ambiguous and misleading news.

Ambiguous headline is unclear meaning between headline and content of the news. The words in the headline need to be chosen carefully. Previous studies used secondly mine class sequential information (CSR) to extract features on headline [21]. They identify word-based, losing sight of sentence position structures and sequential information. Meanwhile misleading headline whose meaning of headline and content are differ from each other [21]. They are trained using Support Vector Machine (SVM) classifier. Please refer Figure 2.

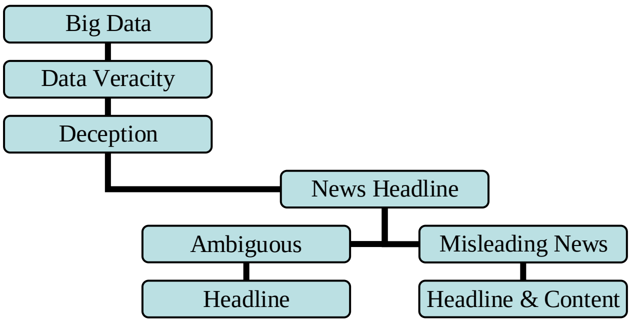


Fig. 2: News headline hierarchy in Big Data Veracity based on Literatures

Deception refers to deliberate misinformation, or an attempt to create a false belief or wrong conclusion. However, it only concern on the quality of information in the disciplines to produce a large amount of textual information (media). Deceptive headlines can create negative impact towards the society especially from the misleading headlines. Among the impact of misleading headlines is the readers tend to be differ from a specific interpretation misleading headlines can also lead to misconceptions and misinformed behavioral to the readers [22].

In this paper, we determine the similarity of news article based on cosine similarity. Bigram features will be included to improve accuracy. Previous research focus on semantic features using LIWC lexicon [15]. They only focus on short text. But we are different because we focus on whole articles to determine deception.

Features is functional or non-functional characteristic of systems [23]. There are a few features that have been highlighted in previous researchers. The features are Absurdity and Humor, Punctuation, Grammar, Body-independent feature, Body-dependent feature, N-gram, Cosine Similarity and Deception Detection measurement.

A. Absurdity and humor features.
Based on [24], there are used named entities such as people, place and location to define unreasonable in a humorous way in text headline.

B. Punctuation.
The punctuations such as comma, period, colon, semi-colon, question marks, exclamation and quotes are included in text. This will give better results to determine deception.

C. Grammar.
They focus on adjectives, adverbs, pronouns, conjunctions and prepositions on grammar. To extract this grammar is using part of speech tagger.

D. Body-independent features.
Researcher focus on headline and extract data to identify misleading headlines [20]. They used co-training approach.

E. Body-dependent features.
Meanwhile this body dependent focus on bodies and headlines [20]. They give name of features likes informality, sentiment, informal gap, sentiment gap and similarity. Approach that they are using is same like body dependent.

F. N-grams features.
N-grams is a contiguous sequence of **n** items from a given sequence of text. Usually many approaches combine the n-gram with other features to detect deception. In our cases we used bigram features.

G. Cosine similarity.
Cosine similarity identifies similarity between two vectors. It calculates the similarity between the headline and content of the news article.

H. Measurement of Deception Detection.
Some initial success in deception detection has created a new wave of applying intelligent technologies to support deception detection [25]. From the previous studies, researchers have mentions about features. This features are important to train data in determine deception detection. For example n-grams, readability and syntactic complexity feature semantics features, shallow and deep syntax features.

Shallow and deep syntax features consist of part of speech (POS) tags and lexicalized production rules derived from Probabilistic Context Free Grammar (PCFG) trees [26]. Semantic features represent the number of words in a sentence belonging to a specific semantic class.

N-grams are contiguous sequence of **n** items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. The **n-grams** typically are collected from a text or speech corpus [27].

Another related approach in [28] use texts features and linguistics cues features to conduct the deception detection. We also can combine those features to enhance the accuracy of deception detection.

There are limitations on deceptive research based on data set available that are accurate. Therefore we are using public dataset for our analysis regarding online news.

Figure 3 below shows the result of deceptive news from the following paper. From this, we can observe that the performance of deception detection is about 60-85%.

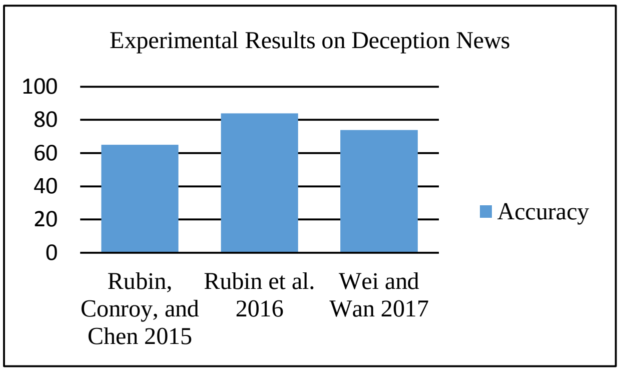


Fig. 3: Experimental Results on Deception News

3. Features Used in Experiment

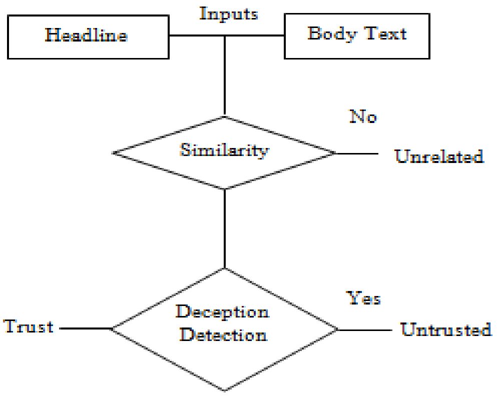
Two types of features are used in this experiment. There are bigram and cosine similarity. Previous researches are not focusing on cosine similarity in detecting deception. This is the reason why we choose to use bigram and cosine similarity.

Bigram: It is called a text with a list of word pairs [29]. For example sentences used: “hi how are you?” and after code into program: Bigram [(‘hi’, ’how’), (‘how’, ‘are’), (‘are’, ‘you’), (‘you’,’?’)]. We used bigram to extract the headline and the content news.

Cosine similarity: It calculates the cosine similarity between counts of word pairs from the headline and content.

4. Proposed Approach

The task for detecting deception for news article is to build a classier to identify trusted or untrusted news article. There consist of 360 labeled article headline and content pairs, which are derived from [8].



Based on the previous studies of headline news observations, we propose a framework to detect deceptive online news based on similarity measure between different news articles. The important and unique part of this framework is the feature-extraction phase in which it will be focusing on the news articles. Then we will use suitable and relevant machine learning approach (SVM) to analyze our result. From the analysis we can predict the deception of misleading online news.
We consider each word in the news articles is very important. Therefore, we are going to use term frequency-inverse document frequency (TF-IDF) method as a part of features extraction in our study. TF-IDF can evaluate how important is a word in document. Mathematically, TF-IDF is expressed as:
TF (t) = (Number of times term t appears in document) / (Total number of terms in the document)

IDF (t) = log_e (Total number of documents) / (Number of document with term t in it).

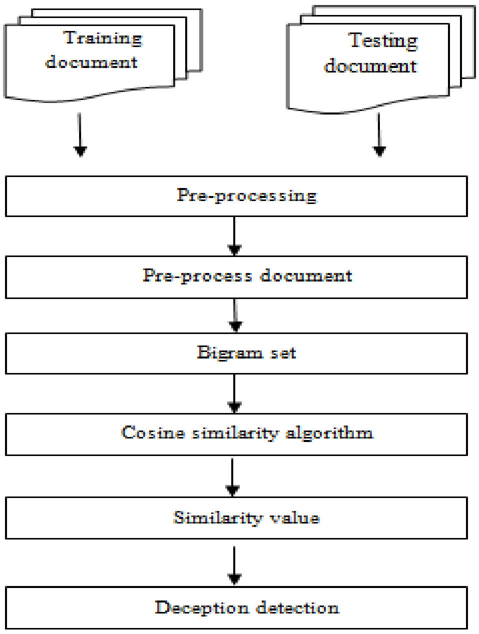


Fig. 4: Steps in the proposed Deception Detection approach

5. Experimental Setup

The similarity between the news articles is measured based on cosine similarity algorithm. The dataset used the collection of 360 news articles that representatives of the scope US and Canadian national newspaper [8]. Data was preprocessed using pandas and the learning and validating process was built with Scikit-learn and NLTK that is implemented by Rubin [8] using Python language.
After the transformation phase, pre-processing will be conducted to normalize data by removing the unrelated data. The tasks involved in this phase are stop word removal and word stemming. Stop removal word is a technique that is used for removing all the stop words, which is a group of word that has no importance in a sentence. Examples of stop words are ‘the’, ‘and’, ‘to’, ‘a’ and more, while stemming is a technique to find the root of words. For example, words such as ‘waited’, ‘waits’ and ‘waiting’ will be reduced to the root word ‘wait.
We also transform the text into lowercase. First step is to define bigram features extraction. We add more contexts to get the frequency of bigram of each news article.
Then pre-process document of news article is performed to get similarity measurement between word pairs. Result of accuracy deception will determine trust or untrusted of news article.

6. Result and Discussion

In order to evaluate our approach, we perform two-class classification where the class label are trust and untrusted. 10 fold-cross-validation are used with support vector machine (SVM).
The accuracy result of bigram is 89% meanwhile accuracy result for similarity measure using cosine similarity is 90% that is good for text deception.

Table 1: Result of deception detection accuracy

Features	Accuracy
Rubin et al (2015)	0.65
Rubin et al (2016)	0.84
Wei & Wan (2017)	0.74
Similarity	**0.90**

The different between our approaches compared existing approach is to detect deception for the whole article using bigram and cosine similarity features.

7. Conclusion

This research paper showed that fake news are intentionally and verifiably false that could misled readers rather than entertain them. Although fake news has been around as long as humankind, it gained increase influence with the printed word and an explosion of influence more recently thanks to the internet and social media platforms. In this paper, we propose an approach for online news deception detection. We included the relation between headlines and its contents similarity into the deception detection approach using SVM. Our proposed approach achieved 90% compared to other deception detection approach.

Acknowledgement

This research was supported by Universiti Putra Malaysia through Putra Grant Scheme-Putra Graduate Initiative (GP-IPS/2016/9478900).

References

[1] N. C. Eembijamil, I. Ishak, and F. Sidi, “Deception detection approach for data veracity in online digital news: Headlines vs

contents," AIP Conf. Proc., vol. 1891, 2017.

[2] Y. R. Tausczik and J. W. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," J. Lang. Soc. Psychol., vol. 29, no. 1, pp. 24–54, 2010.

[3] V. L. Rubin and T. Vashchilko, "Extending information quality assessment methodology: A new veracity/deception dimension and its measures," Proc. Am. Soc. Inf. Sci. Technol., vol. 49, no. 1, pp. 1–6, 2012.

[4] E. Ferrara, "Manipulation and abuse on social media," 2015.

[5] V. Rubin, N. J. Conroy, V. L. Rubin, Y. Chen, and N. J. Conroy, "Deception Detection for News : Three Types of Fakes Deception Detection for News : Three Types of Fakes," no. November, 2015.

[6] V. L. Rubin, N. J. Conroy, and Y. Chen, "Towards News Verification : Deception Detection Methods for News Discourse," no. JANUARY, 2015.

[7] Y. Chen, N. J. Conroy, Y. Chen, N. J. Conroy, and V. L. Rubin, "News in an Online World : The Need for an " Automatic Crap Detector "," no. November, 2015.

[8] V. Rubin, N. J. Conroy, V. L. Rubin, N. J. Conroy, Y. Chen, and S. Cornwell, "Fake News or Truth ? Using Satirical Cues to Detect Potentially Misleading News Fake News or Truth ? Using Satirical Cues to Detect Potentially Misleading News .," no. April, 2016.

[9] H. Allcott and M. Gentzkow, "Social Media and Fake News in the 2016 Election," J. Econ. Perspect., vol. 31, no. 2, pp. 211–236, 2017.

[10] R. M. Entman, "Framing bias: Media in the distribution of power," J. Commun., vol. 57, no. 1, pp. 163–173, 2007.

[11] S. Lee, "Detection of Political Manipulation in Online Communities through Measures of Effort and Collaboration," ACM Trans. Web, vol. 9, no. 3, pp. 1–24, 2015.

[12] "'Fake news' becomes a business model – researchers - The East African." [Online]. Available: http://www.theeastafrican.co.ke/business/Fake-news-a-business-model/2560-4189846-bbkysn/index.html. [Accessed: 29-Apr-2018].

[13] "Identifying Fake News: Use Deception Detection Techniques | Globalytica." [Online]. Available: http://www.globalytica.com/identifying-fake-news-deception-detection-techniques/. [Accessed: 30-Mar-2018].

[14] D. Dor, "On newspaper headlines as relevance optimizers," J. Pragmat., vol. 35, no. 5, pp. 695–721, 2003.

[15] V. Pérez-Rosas and R. Mihalcea, "Experiments in Open Domain Deception Detection," 2013.

[16] E. Ifantidou, "Newspaper headlines and relevance: Ad hoc concepts in ad hoc contexts," J. Pragmat., vol. 41, no. 4, pp. 699–720, 2009.

[17] J. O'Shea, Z. Bandar, and K. Crockett, "A New Benchmark Dataset with Production Methodology for Short Text Semantic Similarity Algorithms," ACM Trans. Speech Lang. Process., vol. 10, no. 4, p. Article No. 19, 2013.

[18] T. Lukoianova and V. L. Rubin, "Veracity roadmap: Is big data objective, truthful and credible?," Adv. Classif. Res. Online, vol. 24, pp. 4–15, 2013.

[19] N. M. Turner, D. G. York, and H. A. Petousis-Harris, "The use and misuse of media headlines: Lessons from the MeNZB??? immunisation campaign," N. Z. Med. J., vol. 122, no. 1291, pp. 22–27, 2009.

[20] W. Wei and X. Wan, "Learning to Identify Ambiguous and Misleading News Headlines," pp. 4172–4178, 2017.

[21] W. Wei and X. Wan, "Learning to Identify Ambiguous and Misleading News Headlines," 2017.

[22] R. Ecker, U.K, Lewandowsky, S., Chang, E.P., Pillai, "The Effects of Subtle Misinformation in News Headlines," Uma ética para quantos?, vol. XXXIII, no. 2, pp. 81–87, 2014.

[23] T. Berger, D. Lettner, J. Rubin, P. Grünbacher, A. Silva, M. Becker, M. Chechik, and K. Czarnecki, What is a feature? 2015.

[24] V. L. Rubin, N. J. Conroy, Y. Chen, and S. Cornwell, "Fake News or Truth ? Using Satirical Cues to Detect Potentially Misleading News .," no. April, pp. 7–17, 2016.

[25] L. Zhou, Y. Shi, D. Zhang, and A. Sears, "Discovering Cues to Error Detection in Speech Recognition Output: A User-Centered Approach," J. Manag. Inf. Syst., vol. 22, no. 4, pp. 237–270, 2006.

[26] S. Petrov and D. Klein, "Improved Inferencing for Unlexicalized Parsing," Proc. NAACL-HLT 2007, no. April, pp. 404–411, 2007.

[27] G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, and L. Chanona-Hernández, "Syntactic N-grams as machine learning features for natural language processing," Expert Syst. Appl., vol. 41, no. 3, pp. 853–860, Feb. 2014.

[28] H. Zhang, Z. Fan, J. Zheng, and Q. Liu, "An improving deception detection method in Computer-Mediated Communication," J. Networks, vol. 7, no. 11, pp. 1811–1816, 2012.

[29] "1. Language Processing and Python." [Online]. Available: https://www.nltk.org/book/ch01.html. [Accessed: 29-Apr-2018].