Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 116 (2017) 3–9

Procedia Computer Science

www.elsevier.com/locate/procedia

2nd International Conference on Computer Science and Computational Intelligence 2017, ICCSCI 2017, 13-14 October 2017, Bali, Indonesia

Keynote Speaker II

Biomedical Engineering Research in the Social Network Analysis Era: Stance Classification for Analysis of Hoax Medical News in Social Media

Mauridhi Hery Purnomo[a], Surya Sumpeno[a], Esther Irawati Setiawan[a,b], Diana Purwitasari[a,c]

[a]*Teknik Komputer, Fakultas Teknologi Elektro, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
[b]*Sekolah Tinggi Teknik Surabaya, Indonesia*
[c]*Teknik Informatika, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
Email: hery@ee.its.ac.id, surya@ee.its.ac.id

Abstract

Biomedical engineering research trend can be healthcare models with unobtrusive smart systems for monitoring vital signs and physical activity. Detecting infant facial cry because of inability to communicate pain, recognizing facial emotion to understand dysfunction mechanisms through micro expression or transform captured human expression with motion device into three-dimensional objects are some of the applied systems. Nowadays, collaborated with biomedical research, mining and analyzing social network can improve public and private health care sectors as well such as research health news shared on social media about pharmaceutical drugs, pandemics, or viral outbreaks. Due to the vast amount of shared news, there is an urgency to select and filter information to prevent the spread of hoax or fake news. We explored in depth some steps to classify hoaxes written as news articles. This discussion also encourages on how technologies of social network analysis could be used to make new kinds improvement in health care sectors. Then close with a description of limitless future possibilities of biomedical engineering research in social media.

Peer-review under responsibility of the scientific committee of the 2nd International Conference on Computer Science and Computational Intelligence 2017.

Keywords: social network analysis, hoax, fake news, sentiment analysis

1877-0509
Peer-review under responsibility of the scientific committee of the 2nd International Conference on Computer Science and Computational Intelligence 2017.
10.1016/j.procs.2017.10.049

1. Introduction

Biomedical engineering research is multidisciplinary works, which commonly related to medical devices used for improving the quality of life. It merges scientific and engineering fields such as nanotechnology, stem cell as regenerative medicine, biomechanics, and biomedical devices for helping diagnostics and therapeutics [1]. As an example, nanoparticles or nanomaterials can be used in diagnostic applications such as gene therapy or bioimaging in the brain. Some of the applied systems within biomedical engineering research [2] are detecting infant facial cry to help to communicate pain [3] and lip feature extracting on the frontal face for speech training of the deaf [4]. Another work is recognizing facial emotion to understand dysfunction mechanisms through micro expression or transform captured human expression with motion device into three-dimensional objects [5]. Those works use image-based data in unobtrusive smart systems for healthcare related models. A recent trend is not only about images but also about textual data in healthcare related models such as fake news or hoax spreading within the power of internet technology.

Fake news or hoax has many definitions such as recurrent issues used as a political weapon, irrelevant truths (post-truths) or intentionally spreading falsehood information (alt-facts) [6]. Alternative facts (alt-facts) are information with no basis in reality while post-truths are defined as beyond the truths or irrelevant information. This discussion has a focus on fake news or hoax in Indonesia [7] in the form of alt-facts [8] especially about medical related issues [9]. Conducted surveys in Indonesia gave perceptions about hoaxes, how they are spread, their topic classifications and their effects on the communities. The medium of communication for hoaxes is varied like image or video such that manipulating non-textual content becomes another problem. Meanwhile, text-based hoaxes are usually spread through social media like Twitter or Facebook and the analysis to recognize them are not limited to the hoax text itself but also on how it is presented, by who, and in what format and context [10]. Hoax text presentation means that the analysis is about linguistic-based features because microblog text like Twitter and email text have different characteristics. The sources who are spreading hoax texts tend to have verifiability issues. Thus in the case of Twitter usage as a medium of communication for hoaxes, analysis of retweeting topology network becomes necessary. Aforementioned hoax analysis can be categorized into two major approaches: linguistic and network. Here, we discuss hoax within online news text that put more weights on the linguistic aspect [11]. Linguistic approaches concern with texts as a bag of equally significant words, syntax structure like noun and verb phrases within the texts, and semantic analysis to recognize any contradictions on other texts with the similar topics of allegedly hoax texts. As network approach, some web-based tools are used to predict and track disease outbreaks through query keyword evaluation [12].

Hoax analysis for the Indonesian language becomes interesting research as well since the government makes it as an important issue [7] [8]. Many unverified but widespread news can hinder the national policies. Hoax texts as bags of words were identified with some classifiers [13] in some conditions such as the experimented topics are general, the dataset comes from undeclared sources, and the training data needs manual label. Several selection features are implemented to filter the important words. Those limitations are understandable because of no Indonesian hoax text datasets available unlike English hoaxes or fake news datasets (http://www.fakenewschallenge.org/). However, those steps are not suitable for analyzing medically related hoax text since the article has convincing content but fabricated facts unless there is a refutation article to verify the false claim (Fig. 1, overall hoax content is showed in the left figure while the disagree texts are in the right figure). Those articles give stance position whether supporting the claims or refuting the hoaxes to evaluate the attitude expressed in the texts. For that reason, in a medical hoax classification problem, stance articles turn out to be essential.

Fake News Challenge (FNC) use stance classification approaches for classifying fake news [14] [15], and their topics are not medically related hoaxes. FNC classify news texts based on headlines concerning their stances of agrees, disagrees, discusses and unrelated. Fig.2 shows article samples with stances of unrelated (top), agree (middle) and discuss (bottom) concerning issues of Christian Bale passes on a Steve Job role. Stance labels need careful evaluation

30-year-old Moscow resident was hospitalized with wounds very intimate nature. As it became known LifeNews, in the hands of doctors, the man complained that his casual acquaintance opoila in the sauna, and then gently held his castration operation.
And actions criminals were executed with surgical precision. … … … (**top, stance: unrelated**)

"Christian Bale will not be starring as Steve Jobs in Aaron Sorkin's upcoming Steve Jobs biopic, according to The Hollywood Reporter. The actor has reportedly decided that he was ""not right for the part,"" deciding to withdraw from the film.
Bale was announced as the star of the film just last week, with Sorkin saying that Christian Bale was ""the best actor in a certain age range"" to play Jobs, and that he had agreed to the part without an audition.… … … (**middle, stance: agree**)

"Actor Christian Bale is in talks to play the leading role in the upcoming Steve Jobs biopic being produced by Sony, reports Variety. Leonardo DiCaprio was originally in talks to play Jobs, but withdrew from negotiations earlier this month.
Penned by Aaron Sorkin, the Steve Jobs biopic was originally set to be directed by ""The Social Network"" director David Fincher, who was said to be in talks with Sony earlier this year, but the director title went to Danny Boyle instead, who also directed the hit movie ""Slumdog Millionaire.""
When David Fincher was still attached to the project, he reportedly was hoping to cast Christian Bale in the lead role due to his undeniable resemblance to the former CEO. Christian Bale is also known for his ability to adapt to roles, shedding and gaining weight as necessary to accurately portray characters.
… … … (**bottom, stance: discuss**)

Fig.2 Stance samples of "Christian Bale passes on role of Steve Jobs, actor reportedly felt he wasn't right for part" (FNC data source)

because headline text and body text can result in different attitudes.

Next section describes some compared methods and experiments as an analysis implementation of medically related hoaxes with stance classification approach using our primary collected small dataset.

Table 1 Comparison methods for hoax or fake news classification

No	References	Hoax Dataset	Experiment Scenarios		Discussions
1	V. L. Rubin, 2017 [10] [11]	(a review)	Linguistic approaches and network approaches (described in Introduction section)		Network approaches are essential for analyzing Twitter or Facebook posts
2	E. Rasywir and A. Purwarianti, 2015 [13] (hoax text classification)	220 collected Indonesian articles, manual labels of hoax and not-hoax, general topics	Preprocessing, weight schemes for feature selection (information gain, mutual information, chi-square, term frequency), and classifiers (Naïve Bayes, SVM, C4.5)		No stemming, probability based features and Naïve Bayes classifier give better results
3	N. Rakholia and S. Bhargava, 2016 [14]	FNC dataset (article dataset + stance headline dataset). Classify the body text with respect claim made in the headline (stance classification)	Neural network approaches for word embedding (deep learning model)	bag-of-words (BoW) features, TensorFlow	Tuning the values of network parameters with greedy approach, i.e. BoW vocabulary size, MultiLayerPerceptron (MLP) hidden layer size, etc
4	B. Riedel, I. Augenstein, G. P. Spithourakis and S. Riedel, 2017 [15]			modified feedforward neural net with regularization	
5	S. Kumar, R. West, and J. Leskovec, 2016. [16]	Wikipedia's page (in English)	Features based on Appearance, Network, Support, and Editor; Classifiers (logistic regression, SVM, random forests)		community help at identifying hoaxes with support & editor features
7	E. Tacchini, et al, 2017. [17]	Facebook posts and their like status with topics of scientific and conspiracy news (in English)	logistic regression, modified boolean label crowdsourcing with harmonic algorithm for the numbers of user see the post and received hoax status of the post		Only need a small number of posts for training compared to full data set.
8	I. Augenstein, et al, 2016. [18]	SemEval 2016 Stance Detection for Twitter (in English)	Capture the stance of a tweet concerning a particular target by learning distributed representations for the tweets. Neural network approaches for word embedding (deep learning model)		conditional encoding for learning target-dependent representations outperform SVM and bag of word vectors
6	W. Ferreira and A. Vlachos, 2016. [19]	article dataset + stance headline dataset (in English)	3-way classification using logistic regression with regularization and alignment features between headlines and claims		number of claims in the dataset give more reliable assessment

2. Methodology

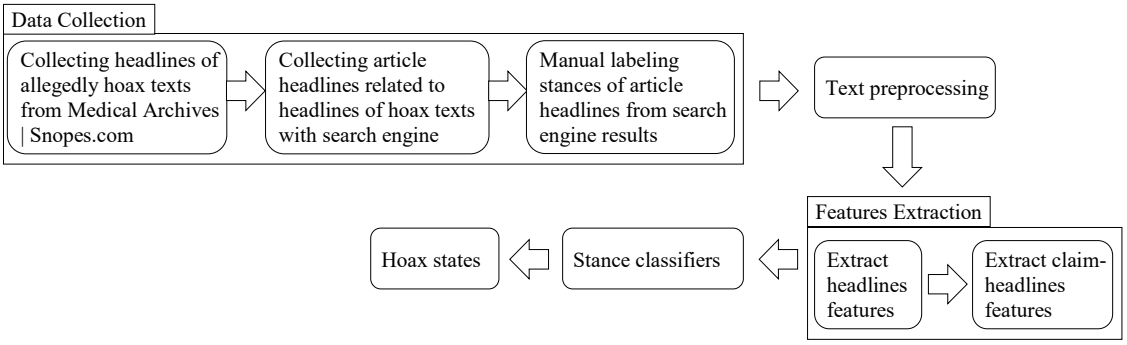

Fig.3 Our experiments for stance detection in hoax analysis with texts of medical claims from Snopes [19]

Some methods are compared for identifying hoaxes with different kind of texts (Table 1) such as Facebook, Twitter, Wikipedia, and article news. Recent researches give more attentions to word-embedding techniques, which need no knowledge about the language of hoax texts [14] [15]. The weight value of term vector is no longer occurrences based but also considering positions of phrase texts like word embedding. However, the text source influences methods of feature selection and classifier. For example, in Wikipedia articles [16] there are appearance features like word counting, content ratio of texts & non-texts, and links within articles. Another kind of feature for Wikipedia articles is network-based coefficients like clustering coefficient to differentiate legitimate and hoax articles. Unique characteristic of Facebook posts create user categories [17], which are user who like hoaxes, user who like non-hoaxes, and user who like both posts. Those categories have an influence in preparing data collection. For Twitter allegedly hoaxes [18], retweeting topology network can be a network feature for hoax analysis to verify the credibility of text source.

This paper showed experiments with steps are directly taken from the approaches of stance classification [19] (Fig. 3). The purpose of our experiments is to show how stance classification implemented in hoax analysis especially with medial contents. The first step of our experiments is collecting data come from medical archives of snopes.com (http://www.snopes.com/category/facts/medical/). The articles are categorized as true, false, and unverified facts. For our Snopes small dataset, we have articles of 19 true claims, 42 false claims, and 17 unverified texts. Then we find 400 related headlines of news texts as stance articles from the search results which have states of **for** when the claim is true, **against** when the claim is false, and **observing** when it merely repeats the claim but uses hedging or vague language. Because of the possibilities for hedging texts, there are two kinds of features: headlines and claim-headlines. Headlines features are common bag-of-words representation from term frequency until distance between root word and refuting word in a sentence that needs a process of analyzing grammatical structure. Whereas, claim-headline features need aligning process between words in a claim with its parallel headline, i.e. Paraphrase Database (PPDB). After aligning words and know their positions, the similarity between vectors of the claim and the headline uses word-embedding results.

3. Result and Discussion

We used the available implementation [19] (https://github.com/willferreira/mscproject) with certain preparation steps for medical hoax dataset. The classifier with headlines features gives varied performances in all stances (Table 2), although claim-headlines features have consistent results in the *against* class (Table 3). It confirms that aligning phrase or words using Paraphrase Database (PPDB) between the claim and the headline is necessary for vague language. Hedging or vague texts are common in the headline instances of *observing* class. Some misclassified data sample are presented in Table 2. The same misclassified texts for headlines are also found in Table 3. In the other hand, accuracy value for the instances of *against* class is consistent between headline feature (Table 2) and claim-headline feature (Table 3). This happens because in the instances of *against* class, the article contents show similar stance with the article headlines.

Table 2 shows the instances of *against* class have better accuracy value compare to others, while the instances of *observing* class have better precision value, and the instances of *for* class have better recall value. This happens because the headline texts are usually short. Therefore, further works on a better understanding of semantic meaning in the article is needed to obtain a better stance. This ongoing research focuses on stance classification, as a preliminary research on hoax analysis. After the stance classification is done accurately, the next step will be on hoax classification.

Table 2 Results with medical hoaxes from Snopes with headlines features

	For	Against	Observing	Accuracy	Precision	Recall	F1	Misclassified Data Sample
For	32	**2**	**1**	0.7432	0.6667	0.9143	0.7711	Does Cuba have a cancer vaccine that has already cured thousands?
	Cuba has a lung cancer vaccine and America wants it (misclassified into **against** class)							
Against	**2**	13	**0**	0.9459	0.8667	0.8667	0.8667	Does drinking cold water after meals cause cancer?
	Drinking cold water after meals can cause cancer (misclassified into **for** class)							
Observing	**14**	**0**	10	0.7973	0.9091	0.4167	0.5714	Sarah Palin wants to invade Ebola?
	The depleted gene pool weighs in on Ebola (misclassified into **for** class)							

Table 3 Results with medical hoaxes from Snopes with claim-headlines features

	For	Against	Observing	Accuracy	Precision	Recall	F1	Misclassified Data Sample
For	28	**1**	**6**	0.7432	0.7000	0.8000	0.7467	Jean Hilliard: Miracle on Ice
	This Girl Was Found Frozen Solid And Almost Dead, Then Something Unexplainable Happened (missclassified into **observing** class)							
Against	**2**	13	**0**	0.9595	0.9286	0.8667	0.8966	Does drinking cold water after meals cause cancer?
	MYTH: Drinking cold water after meals can cause cancer. (misclassified into **for** class)							
Observing	**10**	**0**	14	0.7838	0.7000	0.5833	0.6364	Ebola in Doritos
	4Chan tries to convince people that Doritos are infected with Ebola (misclassified into **for** class)							

4. Conclusion

We have showed that biomedical engineering research are not limited to bioimaging of diagnostics applications for improving the quality of life. Filtering health information in social media such as hoaxes can be associated with biomedical research as well. Because of social media characteristic, linguistic and network features are common approaches in clarifying the truthiness of hoaxes. Understanding the contents of medical hoax or non-hoax texts often uses phrase and word positioning like word embedding. The allegedly hoax texts can be truth claims, false facts, or vague description. For that reason, hoax analysis can be solved as stance classification and data collecting process requires the hoax dataset and the stance dataset. However, text sources such as Twitter or Facebook give another perception for determining hoax analysis. For future works, researches of Indonesian hoaxes still offer many possibilities since analyzing grammatical structure become language barrier.

References

[1] H. Jo, H.-W. Jun, J. Shin and L. SangHoon, Biomedical Engineering: Frontier Research and Converging Technologies, New York: Springer, 2016.

[2] M. H. Purnomo, Y. Kristian, E. Setyati, U. D. Rosiani and E. I. Setiawan, "Limitless possibilities of pervasive biomedical engineering: Directing the implementation of affective computing on automatic health monitoring system," in *8th International Conference on Information Technology and Electrical Engineering (ICITEE)*, Yogyakarta, Indonesia, 2016.

[3] Y. Kristian, H. Takahashi, I. K. E. Purnama, K. Yoshimoto, E. I. Setiawan, E. Hanindito and M. H. Purnomo, "A Novel Approach on Infant Facial Pain Classification using Multi Stage Classifier and Geometrical-Textural Features Combination," *IAENG International Journal of Computer Science,* vol. 44, no. 1, pp. 112-121, 2017.

[4] A. Nasuha, F. Arifin, T. A. Sardjono, H. Takahashi and M. H. Purnomo, "Automatic Lip Reading for Daily Indonesian Words based on Frame Difference and Horizontal-Vertical Image Projection," *Journal of Theoretical and Applied Information Technology,* vol. 95, no. 2, p. 393, 2017.

[5] E. Setyati, M. H. Purnomo, S. Sumpeno and J. Santoso, "Hidden Markov Models based Indonesian Viseme Model for Natural Speech with Affection," *Kursor,* vol. 8, no. 3, pp. 102-122, 2016.

[6] H. Berghel, "Alt-News and Post-Truths in the “Fake News” Era," *IEEE Computer,* vol. April, pp. 110-114, 2017.

[7] Y. Kwok, "Where Memes Could Kill: Indonesia's Worsening Problem of Fake News," 6 January 2017. [Online]. Available: http://time.com/4620419/indonesia-fake-news-ahok-chinese-christian-islam/. [Accessed 19 July 2017].

[8] Masyarakat Telematika Indonesia, "Infografis Hasil Survey MASTEL Tentang Wabah HOAX Nasional," 13 February 2017. [Online]. Available: http://mastel.id/infografis-hasil-survey-mastel-tentang-wabah-hoax-nasional/. [Accessed 26 April 2017].

[9] A. K. Fanani, "Survei menyebutkan hoax terbanyak soal info kesehatan," 1 May 2017. [Online]. Available: http://www.antaranews.com/berita/626813/survei-menyebutkan-hoax-terbanyak-soal-info-kesehatan. [Accessed 5 June 2017].

[10] V. L. Rubin, “Deception Detection and Rumor Debunking for Social Media,” in *The SAGE Handbook of Social Media Research Methods*, London, SAGE, 2017.

[11] N. J. Conroy, V. L. Rubin and Y. Chen, "Automatic Deception Detection: Methods for Finding Fake News," in *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community* , Missouri, USA, 2015.

[12] C. F. Fabricio, “Social networks, web-based tools and diseases: implications for biomedical research,” *Drug Discovery Today,* vol. 18, no. 5/6, pp. 272-281, 2013.

[13] E. Rasywir and A. Purwarianti, "Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin," *Cybermatika,* vol. 3, no. 2, pp. 1-8, 2015.

[14] N. Rakholia and S. Bhargava, "“Is it true?” – Deep Learning for Stance Detection in News," Stanford University, California, USA, 2016.

[15] B. Riedel, I. Augenstein, G. P. Spithourakis and S. Riedel, "A simple but tough-to-beat baseline for the Fake News Challenge stance detection task," Cornell University, New York, USA, 2017.

[16] S. Kumar, R. West and J. Leskovec, "Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes," in *World Wide Web Conference*, Québec, Canada, 2016.

[17] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret and L. de Alfaro, "Some Like it Hoax: Automated Fake News Detection in Social Networks," Cornell University, New York, USA, 2017.

[18] I. Augenstein, T. Rockt¨aschel, A. Vlachos and K. Bontcheva, "Stance Detection with Bidirectional Conditional Encoding," in *Empirical Methods in Natural Language Processing*, Texas, USA, 2016.

[19] W. Ferreira and A. Vlachos, "Emergent: a novel data-set for stance classification," in *The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, California, USA, 2016.