[1]Fake news detection: Network data from social media used to predict fakes

Torstein Granskogen [1] and Jon Atle Gulla[2]

[1] Norwegian University of Science and Technology, Trondheim, Norway
torsteig@stud.ntnu.no
[2] Norwegian University of Science and Technology, Trondheim, Norway
jon.atle.gulla@ntnu.no

Abstract. Fake news has swept through the media world in the last few years, and with that comes a wish to be able to accurately and automatically detect these fakes such that action can be taken against them.

Social network sites are among one of the places where this kind of data are most shared. Using the structure of these sites, we can predict to a high degree if a post is fake or not. We are doing this not by analyzing the contents of the posts, but using the social structure of the site. These social network data mimics the real world where people with similar interests will come together around topics and positions. Using logistic regression and crowd sourcing algorithms, we consolidate previous findings, with prediction accuracy as high as 93 % on datasets consisting from 4200 posts to 15,500. The algorithms show best performance on full datasets.

Keywords: Fake news detection, Social Networks, Contextual Information

1 Introduction

1.1 Problem description

Fake news is a phenomenon that has swept over the world in a massive way the last few years. Suddenly we feel like we are bombarded by news that we cannot know are true or not. To combat this, the scientific community is figuring out ways to automatically detect when a piece of information is reliable or not. In this paper we propose to use a different approach. Our approach is based not on the contents of the news articles, text snippets, tweets etc., but on the traffic and users, and their relations.

As shown in [1] there is a high correlation between the users that actively either comment or like fake articles and stories on Facebook. We want to build on this idea, both by expanding the techniques used by [1], but also by trying to apply it on data that is not as structured as social media. Finally, we want to generate a web-of-trust structure on top of the existing data, that can be used to compute a reliability score for nodes. We

[1] Copyright held by the author(s). NOBIDS 2017

hope that this type of scoring can be used on other actors, such as news agencies, publishers and other important contributors in the information industry.

2 Dataset

The dataset we have chosen to go for is twofold, whereas we have recreated the dataset used in [1] to the best possible match using the same techniques. We are collecting older data, from 2016-07-01 to 2016-12-31. Some of the data is no longer available, and therefore the dataset is not complete, but it contains about one third of the original data. We take this into account when comparing the results to the original ones. The information is volatile, especially the fake parts since Facebook actively removes unwanted information on their site [2].

The data gathered contained the posts from the different sources of scientific and non-scientific sources, together with the likes from those posts, including likes in comments. The likes were concatenation into the post ID, instead of individual comments. The posts were sorted into what community they belonged to, such that a hierarchy of source →post → likes was generated. The identifier for the source was a string of numbers, and each post consisted of the ID sourceID_postID. Following that, the ID of the users was the only information stored per post, no other information about the users were used. The data was manipulated to find the likes from each unique user, but also to find the occurrence of users in the same posts. The datasets were gathered using the Facebook Graph API[3].

2.1 Original dataset

The original dataset consisted of 15,500 posts and 909,236 users, while the one we were able to generate consists of 4286 posts with a total of 158,789 users.

This dataset is a combination of scientific and nonscientific pages. The non-scientific pages are known to publish or embrace fake information, whereas the scientific ones are known to only publish truthful information. This leads to a two-way differentiation, where we have two major nodes that contain the extremes that helps us in differentiating news stories.

2.2 New dataset

In addition to this dataset, we have gathered our own, both to test the same methods as in [1] on a different dataset, but also to check if locale, location or topic have an impact on the results. Locale is the geographical and social affiliation that the users have. The second dataset is divided in the same way as the first, and is comprised of a combination of sources from [4] and [5]. The two sources were needed to get a dataset of similar size and complexity. Not all the sources had a Facebook page, so all of them are not part of the dataset. The complete list of the sources used in both datasets can be found in Table 1.

The new dataset consists of 5943 posts, over 9,5 million likes and 5,6 million unique users. This means that the new dataset consists of less posts, but more users and likes. This is because the sources for the data are mostly from big English or

international mainstream sites, especially the scientific ones, which will then have much greater coverage than the mostly local Italian sites that were used in [1], and containing a bigger spread in locale. This was done to check if a more densely populated dataset with more low-quality users would perform as good as the geographically restricted results as [1].

Table 1. Sources used for datasets.

Original dataset		New dataset	
Scientific	Non-scientific	Scientific	Non-scientific
Scientificast	Scienza di Confine	The Wall Street Journal	Before it's News
Cicap.org	CSSC - Cieli Senza Scie Chimiche	The Economist	InfoWars
Oggiscienza.it	STOP ALLE SCIE CHIMICHE	BBC News	Real News. Right Now.
Queryonline	vaccinibasta	NPR	American Flavor
Gravitazeroeu	Tanker Enemy	ABC News	World Politics Now
COELUM Astronomia	Scie Chimiche	CBS	We Conservative
MedBunker	MES Dittatore Europeo	USA Today	Washington Feed
In Difesa della Sperimentazione Animale	Lo sai	The Guardian	American People Network
Italia Unita per la Scienza	AmbienteBio	NBC	Uspoln
La scienza come non l'avete mai vista	Eco(R)esistenza	The Washington Post	US INFO News
Liberascienza	Curarsialnaturale		Clash Daily
Scienze Naturali	La Resistenza		
Perché vaccino	Radical Bio		
Le Scienze	Fuori da Matrix		
Vera scienza	Graviola Italia		
Scienza in rete	Signoraggio.it		
Galileo, giornale di scienza e problemi globali	Informare Per Resistere		
Scie Chimiche: Informazione Corretta	Sul Nuovo Ordine Mondiale		
Complottismo? No grazie	Avvistamenti e Contatti		
Scienza Live	Umani in Divenire		

2.3 Methodology

The methods used were based on two different algorithms, Logistic Regression(LR) and Harmonic Boolean Label Crowdsourcing(HBLC). LR is a simpler algorithm than HBLC and does not transfer information, whereas HBLC does this. LR considers a set of posts *I* and users *U*, where each post *I* has a set of features $x_{iu}$where x = 1 if a user liked the post and 0 otherwise. The posts are classified based on the users liked them.

The classification is done using a LR model, where each user is given a weight for each user. The summed weight of a post indicated whether it is a hoax or not. The higher the weight, the more likely a post is to be hoax.

HBLC is based on a Boolean label where the label here is True or False. The value is set to be True if the user likes the posts, i.e. gives the post confidence. The dataset is represented by a bipartite graph consisting of the users, the likes and the posts. The harmonic algorithm contains two beta distributions that represents the number of times a user has been seen respectively hoax or non-hoax posts.

HBLC calculates the quality of the post based on these distributions of all the users that have interacted with it, and if the quality is negative it is considered a hoax, and a non-hoax otherwise. Because of the iterative nature of the harmonic algorithm, it can propagate information, such that a hoax user will have an increased value in its hoax beta distribution, and reflected on post beliefs, and consequently infers with the preferences of other, similar users.

A more detailed description of both LR and HBLC can be found in [1].

3 Preliminary results

We have to a been able to recreate the results [1] got using our own version of their dataset with similar results, thereby confirming the findings from [1]. A discussing regarding these results in detail can be found in section 3.2.

Since we were not able to fully recreate the dataset from [1], the results cannot be compared directly. Instead we can use them to test the boundaries for the viability of the different algorithms, and thereby get an indication on how much data is needed for adequate results.

3.1 Dataset results

3.1.1 Original dataset

The results on the smaller dataset we gathered does not impact the results very much, but we see that the smaller the dataset, the more each piece impacts the total score and thus the standard deviation will increase, and the robustness of the results falls.

In addition to these tests, we have done some work on testing other algorithms and how they react to this kind of network data. There is still work to be done to figure out the best parameters using different techniques for this kind of problem, since the data are non-textual and different from what these methods are normally applied on, and to figure out if they are applicable at all.

For the original dataset, we can see that the differences using logistic regression (LR) on the two different versions of the dataset are minor. This is a good indication to LR being a robust algorithm for this kind of data. It performs similarly and predictably on much lower volumes of data. The standard deviation increases, but that is to be expected as the individual posts have a bigger impact in a smaller dataset.

On the other hand, harmonic Boolean label crowdsourcing (HBLC) seems to be more volatile when the size of the dataset decreases. This might be an indication to HBLC needing bigger datasets to perform as good as it did in [1].

3.1.2 New dataset

On the new dataset, we can see that the results are similar the original dataset, which gives a good indication that the algorithms can handle data from different sources.

For LR the results are almost identical to the original dataset. This is an indication that LR is a robust and reliable algorithm. Since the sources were not checked for structural similarities before being collected, this goes to show that if the input data can be divided in non-scientific or scientific groups, LR can be used for good results.

For HBLC, we can see that it performs better than LR overall, but it seems to be more prone to changes when working with smaller datasets. However, on larger datasets, HBLC can predict with very high accuracy whether a post is truthful or not. However, HBLC does not produce as good results on our dataset compared to the one used originally in [1].

Table 2. New dataset, algorithm results.

	One-page-out		Half-pages-out	
	Avg. accuracy	Stdev.	Avg. accuracy	Stdev.
Logistic Regression	0.772	0.288	0.683	0.121
Harmonic BLC	0.939	0.234	0.906	0.102

Table 3. Original dataset compared to our from the same sources.

	One-page-out		Half-pages-out	
	Avg. accuracy	Stdev.	Avg. accuracy	Stdev.
Logistic Regression	**0.794**/ 0.732	**0.303** / 0.363	**0.716** / 0.745	**0.143** / 0.093
Harmonic BLC	**0.992** / 0.978	**0.023** / 0.075	**0.993** / 0.955	**0.002** / 0.062

4 Further work

Going forward, we would like to improve the results we have. This can be done in several ways, and we are going to concentrate on a few of them. First and foremost, we want to look at how further preprocessing of the data will change the results. Since the datasets have a clear majority of users that have a few or just a single like per post, and these users do not contribute much to the result since they have few connections to the rest of the data, removing these or in some way reduce their impact will most likely improve the results.

In addition to this, when using some of the more well-known sites as sources, such as The Wall Street Journal and BBC News, the number of users and data increases rapidly, and the runtime increases even faster. Because of this, a few different approaches can be used. If the system is going to be used in a time sensitive fashion, applying a best-effort algorithm like simulated annealing might help. These kinds of algorithms will give a best possible solution within a given timeframe, and will come closer to the optimal solution the more time it is given to find it. Another way to decrease the complexity is to cluster the users in one way or another. By clustering the users after either closeness to each other or how important they are, the number of operations will be drastically reduced, but some information will be lost to the loss of granularity.

Since the number of usable users are so sparse when dealing with the mainstream sites, this leads to the intersection dataset being really small compared to the total size. An example is the fact that out of over 5.6 million users, only 14 thousand of these have liked posts from both scientific and nonscientific sources. This might be because of the choice of fake news sites, but also indicates that a certain size is needed for a site being viable. To be able to use these algorithms successfully in an industrial setting, we need to be able to either extrapolate the value each user has, or else the intersection dataset will be too small for reliable results.

Because of that, we want to try to apply a web-of-trust, like what was done in [7], on top of the existing results and in that way, try to use that as an early classifier just based on the users. The web will consist of users and the weighted edges between them. Then we can use these weights based on what nodes are already contained in the different posts and then extrapolate and use the social data such as nearest neighbor or clustering to get an indication what these users prefer. Then this score can be used in addition to the one from the algorithms and hopefully give a better indication on whether the post is fake or not.

5 Conclusion

We have shown that logistic regression and harmonic Boolean label crowdsourcing both are viable algorithms on datasets that differs from the original ones that [1] published. In datasets with smaller intersection between the users, both algorithms perform worse, but we hope to remedy this later by further preprocessing of the data. The

algorithms used show robustness in different datasets, one where the number of users compared to pages are small, and another which has more users on a smaller count of pages.

The approach proposed here does also not consider what kind of fake or truthful information is shown, such as whether the fakes are serious fabrications, large-scale hoaxes or humorous fakes, as mentioned in [6].

References

1. Tacchini, E., Ballarin, G., D. Vedova, M. L., Moret, S., de Alfaro, L.: Some Like it Hoax: Automated Fake News Detection in Social Networks. In: Technical Report UCSC-SOE-17-05. School of Engineering University of California, Santa Cruz (2017).
2. CNET Article Mark Zuckerberg on fake news, https://www.cnet.com/news/facebook-fake-news-mark-zuckerberg/, last accessed 2017/11/6.
3. The Facebook Graph API, https://developers.facebook.com/docs/graph-api/, last accessed 2017/11/2.
4. Buzzfeed Political News Data repository, https://github.com/rpitrust/fakenews-data1, last accessed 2017/10/28.
5. PolitiFact's guide to fake news websites, http://www.politifact.com/punditfact/article/2017/apr/20/politifacts-guide-fake-news-websites-and-what-they/, last accessed 2017/10/28.
6. Rubin, V. L., Chen, Y., Conroy, N. J.: Deception Detection for News: Three Types of Fakes. University of Western Ontario, London, Ontario (2015).
7. Tavakolifard, M., Almeroth, K. C., Gulla, J. A.: Does Social Contact Matter? Modelling the Hidden Web of Trust Underlying Twitter*. In: WWW '13 Proceedings of the 22nd International Conference on World Wide Web, p. 981-988. Norwegian University of Science and Technology, Trondheim, Norway and University of California at Santa Barbara, Santa Barbara, USA (2013).