Using Big Data for Fact Checking

Karthik Vegi
Indiana University Bloomington
2619 East 2nd Street, Apt 11
Bloomington, IN 47401, USA
kvegi@iu.com

ABSTRACT

In this data age, the sheer volume of data makes it impossible to know what is truth and what is not. Politicians are often misconstruing facts to improve their candidacy. Scientists and advertisers are making false claims to gain business advantage. The more the false claims penetrate into the internet, especially social media, the more chances are that it is believed to be true. We show how Big Data techniques can be used to spot fake news, false claims made by politicians, advertisers, and scientists.

KEYWORDS

i523, hid231, big data, veracity, fact check, data accuracy

1 INTRODUCTION

Big Data is playing a crucial role in building a smarter planet. Each and every action that we take leaves a digital footprint. Big Data is lending a great helping hand to crunch this data and make smarter decisions. "*Big Data* is at the heart of the smart revolution. It is already completely transforming the way we live, find love, cure cancer, conduct science, improve performance, run cities, and countries and operate business" [6].

Analyzing data in this digital era where data can come from multiple sources involves reading data from different systems in different formats with different contextual meanings. The data extracted from multiple systems can often contradict each other. It could be biased towards a business or a particular entity. Multiple sources also mean conflicting and outdated information which makes it highly inaccurate [1].

Validation of facts became a major issue with the recent U.S. election of 2016 where the candidates from both the democratic and republic parties used a lot of factual statements in the debates to put their candidacy and party in a better position. These factual statements if not validated might give a false edge to the party thus having an effect on the entire nation. While some people take these statements with a pinch of salt, a large set of population often believes it to be true and end up voting for party purely based on the claims made by the respective candidates [2]. With so many data sources like social media, print media, and the internet, it is not easy to validate and spot fake news. We need to take the help of the technological advances in *Big Data* and *Artificial Intelligence* to handle this problem.

Data inconsistencies with respect to the sources, interpreting the data out of the context, obsolete data and data that is highly modified from the original are all data veracity problems [1]. "Fake news and fact-checking is clearly a data veracity problem. Veracity refers to several quality dimensions related to repairing data inconsistencies and fixing other data quality problems such as duplicates, missing or incomplete data" [1].

2 FACT CHECKING AS A BIG DATA PROBLEM

Often veracity is not just about data quality, it is about data understandability. Fake news is understandable and we can make great sense out of it by careful analysis [2]. The data veracity problem in *Big Data* meets the fake news problem at the juncture called *Misinformation Dynamics* where the emphasis is not just on the inaccuracy because of an accident, but on the data quality as a whole [2]. Fake news is often intentional and moreover, it is not static but dynamic [2].

One straightforward way to understand and account for the reliability of the sources is to formulate a voting algorithm that labels the source system in which the data item resides thus evaluating the accuracy of the data [3]. The problem with this approach is that it is too simple and also it doesn't take into account the other factors such as the data lineage of each source [3]. This means that if multiple sources of the data are all derived from another source which is inaccurate, we are wrongly labeling the data source [3].

The social networking giants like Facebook and Twitter faced this problem and a lot of fingers were pointed at them for acting as a medium for spreading fake news. Facebook took the initiative to tackle the problem head-on by implementing an option where the users can flag the story as either true or false [2]. The more false votes s story garners, the less likely it appears on the news feed along with a warning message to the users mentioning that a lot of users have reported the story as false [2]. The problem with this approach is that we are giving people a chance to alter the truth, also making everyone believe that anything that is not flagged is true which might not always be the case [2].

In order to solve this problem in a more efficient way, we could combine *Big Data* and *Artificial Intelligence* techniques that eradicate the possible human-generated errors [2]. *Google* came up with a new method of scoring web pages based on the accuracy of the facts in which the algorithm assigns documents a trust score taking the context into account, which feeds the overall scoring to determine the search rank without solely relying on the links [2].

While the social networking giants have a huge role to play in identifying fake news, each individual should take personal responsibility to check the validity of the data using online tools at their disposal rather than believing it blindly.

3 BIG DATA TECHNIQUES FOR FACT CHECKING

3.1 Recommendation Based Approaches

Recommendation based approaches take the help of the community to determine the accuracy and quality of the sources [1]. The reputation of the sources increases as more people agree that the source is reliable [1]. These methods clearly have their shortcomings as people can be influenced by third-party agencies to improve the trustworthiness of certain sources [1].

3.2 Content Based Approaches

Content based approaches work by coming up with a score to compute the trust of a source and validates the belief of the claims it is making to generate a belief score [1]. The trustworthiness of the source now becomes a function of the trust score and the belief score [1]. This is not a one time process but the function runs over and over to continuously update the source quality [1]. A few probabilistic methods are added to this function to improve the accuracy of the score which extends the algorithm beyond just trust and belief [1].

In one such application, the truth discovery problem is transformed into a probabilistic inference model [7]. An iterative algorithm is proposed which computes the posterior distribution of all the values of the sources and finds the one with the maximum probability [7]. The model derives all the possible values reported by the sources and the conflicting values in the data streams and then calculates a score [7].

Figure 1 illustrates the content based approach for truth discovery in data streams. As there can be heterogeneous sources, first a semantic mapping is employed for the values provided by various sources such that the values for truth discovery are consistent [7]. Taking an example, the weather conditions that imply the same meaning such as *rainy* and *wet* are considered to be the same in truth discovery [7]. The same way, *partly sunny* and *cloudy* are considered as *clear* weather condition [7].

"At each time i, the system collects a set of conflicting values for entity v as $V = \{v1, v2, ..., vk\}$ from multiple data sources. Next, the system resolves the conflicts and discovers the true value v in V based on the current data uncertainty and source quality. Then, the system updates the data uncertainty and source quality based on the inferred value v and conflicting values V" [7].

[Figure 1 about here.]

3.3 Evidence Based Approaches

Evidence based approaches augment the content based approaches by relying on evidence, context and prior knowledge about the data sources [1]. Data provenance information may be used in truth discovery computation, as well as external information about the context, the sources, the data or user network [1]. This involves checking the dynamics of information in the network and recomputing the truth discovery accordingly [1]. Not every industry has a separate budget for research which makes evidence based approaches a viable option only for big organizations.

4 REAL-TIME FACT CHECKING

In this digital age, fact checking makes more sense when it is done in real time. Politicians and media houses use inaccurate facts and make claims and get away with them in real time, but the new fact-checking tools can often expose claims that are invalid and inaccurate [4]. The number of active fact-checking websites has been growing immensely, from about 44 in 2014 to about 114 currently [4].

The delay window between the time when a claim is made and the time when the claim is checked for truth has to be as less as possible as fact checking often takes longer time than traditional journalism [4]. This gives enough time for the politicians and other people to make a claim and get away with it [4].

4.1 ClaimBuster

ClaimBuster is a system that is built for real-time fact checking using the techniques of natural language processing and machine learning combined with database queries [5]. Although the complete system is still in works, some components of the system are already in use [5].

Figure 2 shows the system architecture of *ClaimBuster*. The *claim monitor* integrates the various data sources and feeds them into the system [5]. The *claim spotter* picks the claims that could potentially be checked for accuracy and reads in the relevant text from the data sources [5]. The *claim matcher* finds all the matching data sources that which mention the same claim in different ways [5]. The *claim checker* verifies them against external information from the internet to compute the accuracy of the claim [5]. The *fact-checker reporter* validates the claims against the facts gathered from *claim matcher* and compiles the accuracy report to publish them through sources like a website, a twitter account, or a slack-bot [5].

In this way, the *ClaimBuster* gives every fact a score between 0 to 1, where a score closer to 1 is more accurate [5]. The model was well trained by using thousands of actual data from general election debates that has been manually fact-checked by humans [5]. The accuracy of the model was measured between 74 and 79 [5]. The system was put to use in real-time for the 2016 U.S. presidential election debates and the results showed a high match between *ClaimBuster* and journalists who checked for the accuracy of the claims [5].

[Figure 2 about here.]

5 CONCLUSION

Big Data coupled with *Artificial Intelligence* and *Machine Learning* can tackle the fact checking problem more efficiently. Rather than working in silos, the social networking giants and the search engine giants should work together with researchers to improve the existing system. This ensures that there are no loose ends with respect to the accuracy of the data. This is important because there is always a disconnect between data sources and not everybody has control and access to data that somebody else owns.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and the teaching assistants for their support and suggestions in writing this paper.

REFERENCES

[1] Laure Berti-ftquille and Javier Borge-Holthoefer. 2015. *Veracity of Data: From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics*. Morgan & Claypool Publishers, Qatar.

[2] Forbes. 2017. Fake News - Big Data And Artificial Intelligence To The Rescue. Webpage. (Jan. 2017). https://www.forbes.com/sites/jasonbloomberg/2017/01/08/fake-news-big-data-and-artificial-intelligence-to-the-rescue/#69e474df4a30

[3] Forbes. 2017. Fake News: How Big Data And AI Can Help. Webpage. (March 2017). https://www.forbes.com/sites/bernardmarr/2017/03/01/fake-news-how-big-data-and-ai-can-help/2/#7ea468b92039

[4] Naeemul Hassan, Bill Adair, James T. Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015. The Quest to Automate Fact-Checking. In *Proceedings of the 2015 Computation + Journalism Symposium*. Brown Institute of Media Innovation, New York, NY, USA.

[5] Naeemul Hassan, Anil Kumar Nayak, Vikas Sable, Chengkai Li, Mark Tremayne, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, and Aaditya Kulkarni. 2017. ClaimBuster: the first-ever end-to-end fact-checking system, In Proceedings of the VLDB Endowment. *Very Large Database Endowment* 10, 1945–1948.

[6] Bernard Marr. 2015. *Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance*. Wiley, Chichester.

[7] Zhou Zhao, James Cheng, and Wilfred Ng. 2014. Truth Discovery in Data Streams: A Single-Pass Probabilistic Approach. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)*. ACM, New York, NY, USA, 1589–1598. https://doi.org/10.1145/2661829.2661892

List of Figures

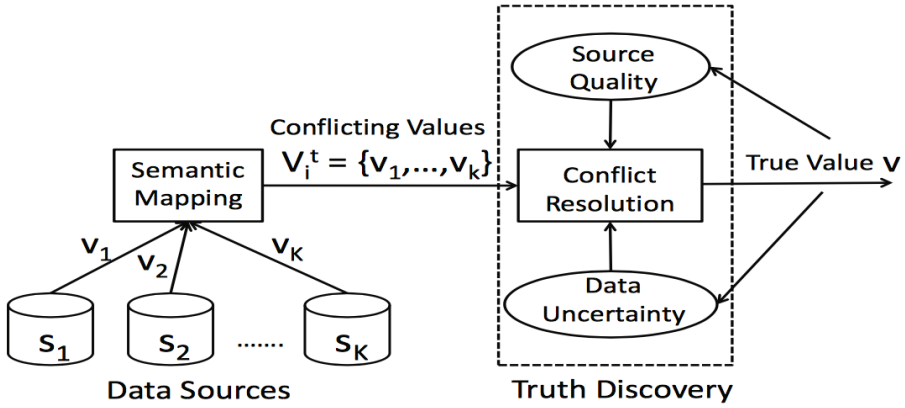

Figure 1: Truth Discovery In Data Streams [7]

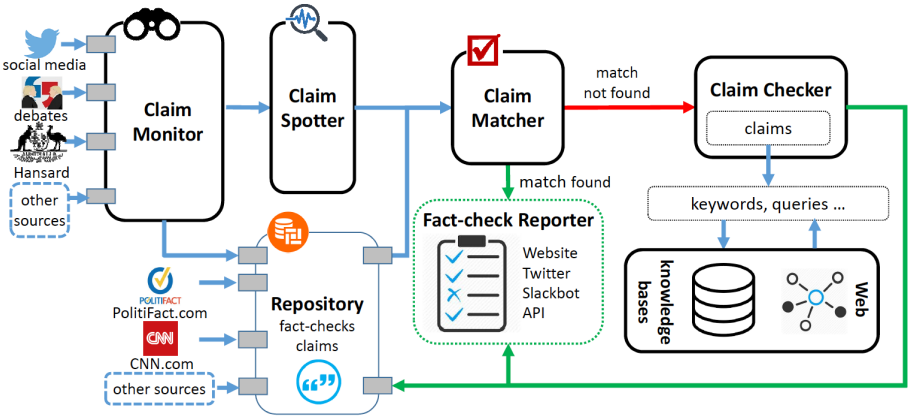

Figure 2: System architecture of ClaimBuster [5]