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

Big Data is no more the elephant in the room it once used to be. Since John Mashey coined the term in 1998, it has come a long way. It is often described as the three Vs: Volume, Velocity, and Variety of the data. Of late, a new dimension, *Veracity* has been gaining importance which describes the quality and accuracy of the data. We show how Big Data can be used to spot fake news, bad data used 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 to 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 [5]"

"Large scale searches and analyses over multiple sources involves extracting data from highly heterogeneous structures, semantics, and qualities. One of the fundamental issues is that the extracted information can be biased, noisy, outdated, incorrect, misleading, and thus unreliable. To add to the problem, available data sources can provide conflicting information, leaving the users in doubt with respect to the accuracy [3]."

"The impact of fake news on the recent election has focused public attention on this multi-tentacled and growing problem. Vast swaths of the population fall prey to such misinformation, while others struggle to discern unbiased truth from the morass of lies and distortions that surrounds us [1]." With so many data sources like media, internet, newspaper, and many more, it is not easy to spot fake news and fack check the data. We need to take the help of the technological advances like Big Data and Artificial Intelligence to handle this problem.

"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. Data veracity can be attributed to the following:

Ambiguity: Data can be inconsistent from one source to another, leading to misinterpretation.

Staleness: The data is obsolete and no longer relevant.

Falsification: False or distorted information can be intentionally propagated by one source or a coalition of sources. Information can be manipulated or presented selectively to influence the audience and encourage and particular conclusion [3]."

2 FACT CHECKING AS A BIG DATA PROBLEM

Often veracity is not just about data quality, it is about data *understandability.* But fake news is understandable and we can make great sense out of it by careful analysis. We should therefore strive to achieve *truthfulness* [1]. "Misinformation dynamics, in fact, is where the big data concept of data veracity and the problem of fake news connect. We are not simply talking about the accidental inaccuracies that make up the bulk of enterprise data quality efforts. On the contrary, fake news is intentional misinformation, and furthermore, it is dynamic [1]."

"A common strategy to evaluate the reliability of the sources is to take advantage of data redundancy, and rely on majority voting heuristic, which simply assigns a true label to data that are claimed by the majority of the sources. But this strategy is known to be error-prone, because it counts all the sources equally and does not consider source dependence [2]."

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 false. The more false votes it garners, the less likely it is for it to appear on the news feed. It also displays a warning to the users mentioning that a lot of users have reported the story as false. But the problem here is that we are giving people a chance to alter truth. It also makes everyone believe that anything that is not flagged is true which might not always be the case [1] .

"To solve these problems, a combination of big data and AI methodologies are being developed that rely less on human-generated input, which can be swayed by opinion or a lack of facts. Google published a paper in 2015 about a new method of scoring web pages based on the accuracy of the facts presented. The algorithm assigns documents a trust score, which would then presumably be used as part of Google's overall scoring to determine search rank. The technology is important, because it is attempting to understand a page's context without the use of third-party signals, like links [2]."

"The news media and social media cannot be solely responsible for preventing fake news. Each one of us have equal responsibility to discern the accuracy. Tools already exist that can help individual users spot fake news sites. *Hoaxy* is an online tool that helps people visualize the spread of claims and fact-checking online, and is available to anyone to use. Many Chrome extensions have been created that can alert and help filter fake news. Even popular websites like

Snopes and *FactCheck.org* can help identify the most egregious fake stories [1]."

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. The reputation of the sources increases as more people agree that the source is reliable. These methods clearly have their shortcomings as people can be influenced by third party agencies to improve the trustworthiness of certain sources [3].

3.2 Content Based Approaches

"Content based approaches work by computing a trustworthiness score of a source as a function of the belief in its claims, and then the belief score of each claimed data as a function of trustworthiness of the sources asserting it [3]." The source quality is initialized and iteratively updated based on the content belief. Various probabilistic methods have been used to tackle other aspects beyond trustworthiness and data belief [3].

In one such methodology, the truth discovery problem is transformed into a probabilistic inference model. 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. 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]. "For example, the meaning of the weather conditions ***rainy*** and ***wet*** are considered to be the same in weather forecast truth discovery. Also ***partly sunny*** and ***mostly cloudy*** are grouped and considered to be the same as ***clear*** [7]."

"At each time t, the system collects a set of conflicting values for entity i as $\boldsymbol{V} = \{v1, v2, ..., vk\}$ 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 $\mathbf{v}$ 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 priori knowledge about the data sources [3]. 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 [3]. This involves checking the dynamics of information in the network and recomputing the truth discovery accordingly [3]. "The problem with evidence based practice is that outside of areas like health care and aviation is that most people in organizations do not care about having research evidence for almost anything they do. That does not mean they are not interesting in research but they are just not that interested in using the research to change how they do things [6]"

4 AUTOMATING FACT CHECKING

In this digital age, fact checking makes more sense when it is done in real time. "Politicians and media figures make claims about *facts* all the time, but the new army of fact-checkers can often expose claims that are false, exaggerated or half-truths. The number of active fact-checking websites has grown from 44 a year ago to 64 in 2015, according the Duke Reportersfis Lab [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. Fact checking takes longer time than traditional journalism. This gives enough time for the politicians and other people to make a claim and get away with it [4]

4.1 Computational Challenges

4.1.1 ***Finding claims to check:***

This constitutes converting raw data to natural language and extracting contextual information such as speaker, time, and occasion [4].

4.1.2 ***Getting data to check claims:***

This involves evaluating the quality and completeness of sources and mapping them back to the data sources. Integrating multiple sources and cleansing data is an integral part of this step [4].

4.2 Claimbuster

Claimbuster is an online tool to check for facts in real time. "For every sentence spoken by the participants of a presidential debate, Claimbuster determines whether the sentence has a factual claim and whether its truthfulness is important to the public. The calculation is based on machine learning models built from thousands of sentences from past debates labeled by humans. The ranking scores help journalists prioritize their efforts in assessing the veracity of claims. Claimbuster can be expanded to other discourses such as interviews and speeches and also adapted for use with social media [4]."

Claimbuster makes use of a supervised learning approach and breaks the sentences into three categories namely *Non Factual Sentences*, *Unimportant Factual Sentences*, and *Check-worthy Factual Sentences* [4]. "Given a sentence, the objective of Claimbuster is to derive a score that reflects the degree by which the sentence belongs to *Check-worthy Factual Sentences*. Many widely-used classification methods support ranking naturally. For instance, consider a Support Vector Machine (SVM). *Check-worthy Factual Sentences* are treated as positive examples and both *Non Factual Sentences* and *Unimportant Factual Sentences* as negative examples. SVM finds a decision boundary between the two types of training examples and calculates the posterior probability using a decision function. The probability scores of all sentences are used to rank them. This clearly will help the journalists and fact checkers to free up time to focus on more important things like reporting and writing [4]."

5 CONCLUSION

Big Data coupled with Artificial Intelligence and Machine Learning can tackle the fact checking problem. Rather than working in silos, the social networking giants and the search engine giant could work together with researchers to come up with a more effective solution. This ensures that there are no loose ends with respect to the accuracy of the data. This is important because there is a disconnect between data sources at times 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] 2017. Fake news: Big Data and AI 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

[2] 2017. Fake news: How Big Data 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

[3] Laure Berti-ftquille and Javier Borge-Holthoefer. 2016. *Veracity of Data*. Morgan & Claypool.

[4] Naeemul Hassan, Bill Adair, James Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015. The Quest to Automate Fact-Checking. (10 2015).

[5] Bernard Marr. 2015. *Big Data*. John Wiley & Sons Ltd.

[6] Oxford. 2018. Evidence based practice problem. Webpage. (2018). https://www.oxford-review.com/blog-research-problem-evidence-based/

[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

6 BIBTEX ISSUES

Warning–no key, author in www-forbes2

Warning–no author, editor, organization, or key in www-forbes2

Warning–to sort, need author or key in www-forbes2

Warning–no key, author in www-forbes1

Warning–no author, editor, organization, or key in www-forbes1

Warning–to sort, need author or key in www-forbes1

Warning–no key, author in www-forbes2

Warning–no key, author in www-forbes2

Warning–no key, author in www-forbes1

Warning–no key, author in www-forbes2

Warning–no author, editor, organization, or key in www-forbes2

Warning–empty author in www-forbes2

Warning–no key, author in www-forbes1

Warning–no author, editor, organization, or key in www-forbes1

Warning–empty author in www-forbes1

Warning–empty address in Berti-Equille2016

Warning–no journal in Hassan2015

Warning–no number and no volume in Hassan2015

Warning–page numbers missing in both pages and numpages fields in Hassan2015

Warning–can't use both author and editor fields in Marr2015

Warning–empty address in Marr2015

Warning–numpages field, but no articleno or eid field, in Zhao2014

(There were 22 warnings)

7 ISSUES

DONE:

Example of done item: Once you fix an item, change TODO to DONE

7.1 Writing Errors

this is a collection of quotations, not a paper

7.2 Citation Issues and Plagiarism

Need to paraphrase long quotations (whole sentences or longer)

Need to quote directly cited material. Are you sure you have quoted all of them?

The citation mark should not be in the beginning of the sentence or paragraph, but in the end, before the period mark and after a quotation mark. example: "a library called Message Passing Interface(MPI)" [7].

List of Figures

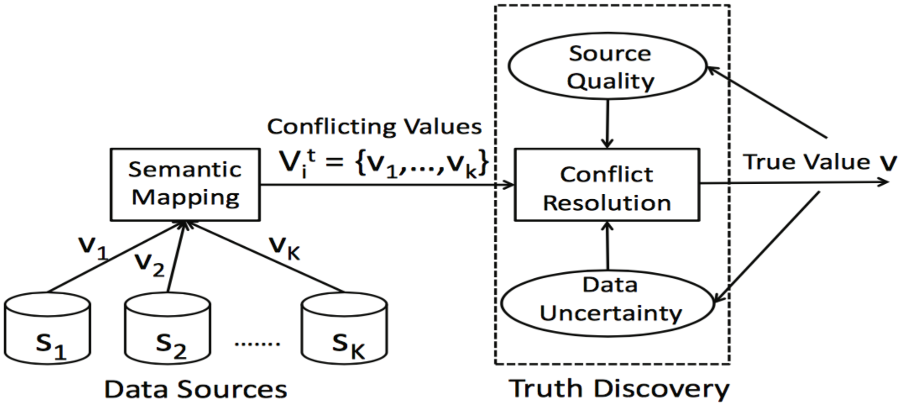

Figure 1: Truth Discovery In Data Streams [7]