Written Representation 115

Name: Tisane Labs Pte. Ltd.

Received: 5 Mar 2018

2018 Parliament Select Committee on Online Falsehoods: Tisane Labs
By Vadim Berman (Tisane Labs Pte. Ltd.)

Introduction

This purpose of this paper is to survey the technical capabilities and limitations of both the dissemination and the automatic detection of online falsehoods.

Automatic Detection

Truth

The core question is the feasibility of finding out truth from a text alone. Software engineering, whether it's machine learning, any other brand of artificial intelligence, or the plain old Boolean logic-based software, is not magic; it has its limitations. As of today, most of AI works by trying to approximate human judgement.

Humans are unable to confidently conclude the authenticity of a statement without so-called expert knowledge. Let's use a simple example. Suppose, the author of this document claims that he is 256 years old. An average human will be able to figure out that it is not a true statement relying on the knowledge that humans don't live that long. If the author's claim is that he is 81 years old, it is more believable but most people would still not trust the claim because the share of people in this age bracket in the workforce is low, and even lower in the software engineering industry. If, however, the author claims he is 45, is he lying or telling the truth? There is no way of telling without learning more about the author.

Claims of Fully Automatic Detection of Fake News

Every now and then, emerges an academic paper or a vendor's claim of being able to detect fake news automatically. In brief, none of them can provide a real comprehensive solution simply due to the fact that the actual truth cannot be verified by the means of linguistic analysis.

These offerings broadly fall in two categories:

- Machine learning solutions that simply derive their conclusions based on similarity of texts. If a text is similar enough to a text marked as "true", it is deemed true. Otherwise, false. Needless to say, this approach is extremely naïve and can only work with demos and toy data. The more primitive tools that employ so-called "bag of words" approach, may be fooled by adding modifiers like "not" which reverse the meaning of a claim but will not impact the similarity score.
- Deep linguistic analysis solutions that weigh the number of allegations vs. the number of factual statements in an article. This approach may evaluate the standards of journalism but, yet again, has no way of determining whether a particular claim is based on an actual fact. The author may be an articulate

and an eloquent liar, or tell the truth in a way that does not sound very believable.

Feasible Solutions
While fully automatic detection of falsehoods is not feasible, it is very much within the current level of technological development to **detect texts and posts on controversial topics** that stir the public opinion.

Combined with human supervision and the analysis of engagement of the readers, it is possible to build a workflow that detects the spread of online falsehoods and allows to react in a timely manner without breaking the bank and false positives. It will also address the side issue of “half-truths” or malicious take on factual information.

Dissemination
The dissemination of falsehoods has been employed by both private interests and state actors since time immemorial. The recent technological advances, however, boosted abilities in several areas:

- **Measurable feedback**. The proliferation of social media and measurable reactions (e.g. Facebook Like button) created a reliable way to index the public engagement, perhaps more reliable than public opinion polls.
- **Speed of dissemination**. Social media channels optimise delivery of news by demand, using subscription models and automatically promoting popular posts. This allows viral spread of content that the public finds interesting and relevant, much faster than traditional media ever dreamed of.
- **Accessibility**. It costs virtually nothing to publish a post that would rival the influence of the traditional media. This removes barriers to participations of actors that in the past would not be able to afford exercise influence on the public opinion. All it takes now is determination, creativity, and a clear objective.