Natural Language Processing for Fact-Checking and Claim Assessment

Intermediate report of project advancement

Othman EL HOUFI
Dimitris KOTZINOS

MSc Research in Data Science & Machine Learning

December 13, 2021

Abstract

As false information and fake news are propagating though out the internet and social networks, the need of fact-checking operations becomes necessary in order to maintain a truthful digital environment where general information can be reliably exploited whether in politics, finance and other domains. The need of this online claim assessment comes from the fact that fake news and false information can have a big negative impact on politics, economy (2016 USA Elections) and public health (COVID-19).
A number of solutions have been proposed to deal with this problem and limit the spread of false information, both manual and automatic. Of course the manual approaches done on websites such as *PolitiFact.com*, *FactCheck.org* and *Snopes.com* don't construct a viable solution for the long term as the speed and scale of information propagation increase exponentially rendering this manual fact-checking operation where human fact-checkers can't scale up at the same rate limited and incapable of solving the problem.
Here, we present our contribution in this regard: an automated solution for fact-checking using Wikipedia's articles for claim verification. The algorithm uses NLP techniques in order to extract the so-called claim from the user input, then, using Wikipedia's API, it retrieves all the relevant articles and assesses with a degree of confidence if the claim is true, false or unable to decide due to lack of information showing evidence (sentences in articles) and probabilities for each resulted case.

Keywords: Natural Language Processing, Wikipedia, Information retrieval, Text processing, Natural Language Inferencing, Fact-Checking, Document retrieval, Sentence retrieval, Fake-news.

Contents

Chapter 1

Introduction

1.1 Project Context

From a social and psychological perspective, humans have been proven irrational and vulnerable when differentiating between truth and false news (typical accuracy ranges between 55% and 58%), thus fake news obtain public trust relatively easier than truthful news because individuals tend to trust fake news after repeated exposure (*Validity effect*), or if it confirms their pre-existing beliefs (*Confirmation bias*), or simply due to the obligation of participating socially and proving a social identity (*Peer pressure*). The social sciences are still trying to comprehend the biological motivations that makes fake news more appealing to humans.

On the other hand, the growth of social media platforms resulted in a huge acceleration of news spreading whether true or false. As of Aug. 2017, 67% of Americans get their news from social media. These platforms even give the user the right to share, forward, vote and participate to online discussions. All of this made the problem of fake news spreading more and more dangerous, our economies for example, are not robust to the spread of falsity, false rumors have affected stock prices and the motivations for large-scale investments, as we witnessed after a false tweet claimed that Barack Obama was injured in an explosion which caused $130 billion drop in stock value. Another recent example is related to public health where rumors about COVID-19 vaccines and drug companies influenced people in their decision on getting vaccinated.

That being said, is there a way to monitor the spread of fake news through social media? Or more specifically, how can we differentiate between fake news and truthful news, and at what level of confidence can we do that?

From a computer engineering perspective, different approaches were studied:

- **Knowledge-based Fake News Detection:** a method aims to assess news authenticity by comparing the knowledge extracted from to-be verified news content with known facts, also called fact-checking.
- **Style-based Fake News Detection:** focuses on the style of writing, i.e. the form of text rather than its meaning.
- **Propagation-based Fake News Detection:** a principled way to characterize and understand hierarchical propagation network features. We perform a statistical comparative analysis over these features, including micro-level and macro-level, of fake news and true news.
- **Credibility-based Fake News Detection:** the information about authors of news articles can indicate news credibility and help detect fake news.

In this project we will focus on the method of **Knowledge-based Fake News Detection** also called **Fact-Checking**. The goal is not to implement an algorithm that scans social networks for real time fake news detection, but rather we will create a model that can assess with a degree of confidence the truthfulness or falseness of a claim given by a user as an input by exploiting Wikipedia's articles as a source of true knowledge and export evidence that validates or refutes the subjected claim.

1.2 Use case scenario

Suppose that while browsing the internet or talking to people you come across a claim that says *"The former U.S president John F. Kennedy died in September 22, 1963"*, as it is a general truth and not a relative truth it should be easier to verify the validity of this claim as well as find evidence that proves it.
Using the platform we will create, you can simply write the claim you like to verify with no regards to a specific linguistic rule, the model will extract relevant articles from Wikipedia using an API, then it retrieves sentences relative to your claim and apply a comparison in order to assess if the claim is True, False, or Not Enough Information as well as giving a percentage of confidence and evidence of the results that were processed straight from Wikipedia's database.
Combing back to our example, the model should return that *"John F. Kennedy died in November 22, 1963"* so the input claim is false.

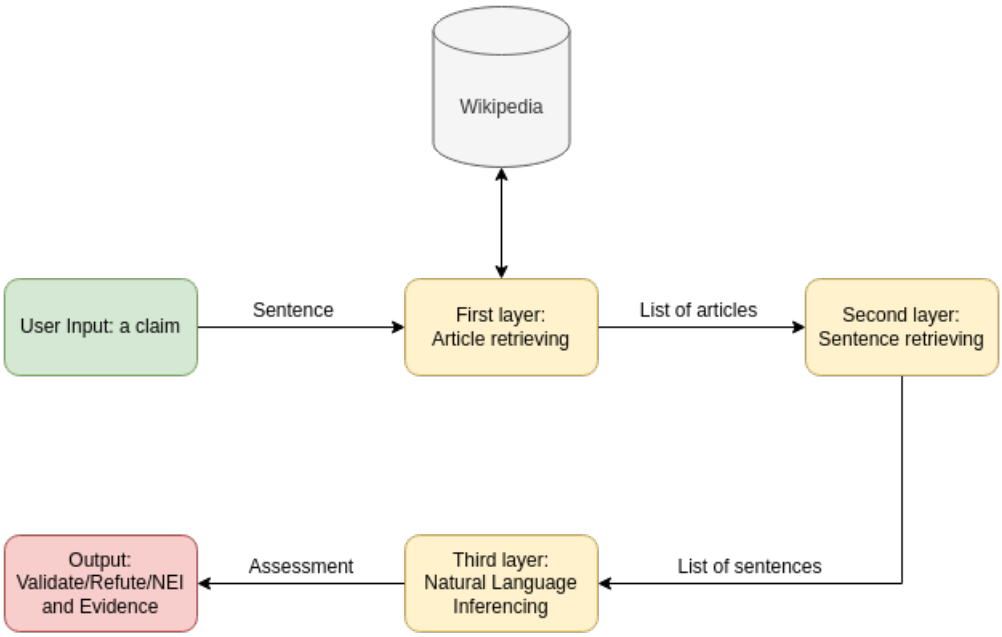

Figure 1.1: General Model Processing Pipeline

Chapter 2

Identified challenges and solutions

2.1 Fact-Checking challenges

In order for the model to work each layer/part of the system must answer to a specific task, the combination of results made by each layer constructs a robust model that is able with a degree of confidence to fact-check a claim as well as present evidence of the assessment. Although for each layer to work as intended we must find solutions to these challenges:

- **Claims Spotting:** the model must be robust to linguistic changes, we must deal with different phrasing for the same or similar claims. For every "reasonable" input we must extract the target claim.
- **Articles Retrieving:** as Wikipedia holds millions of articles, the model must look for a limited number of relative articles to the input claim with an order of degree of correlation.
- **Time Recording:** relative articles in Wikipedia can be outdated, for example Britain belongs to the European Union is an outdated knowledge. The model must be sensitive to the timestamps of articles.
- **Sentences Retrieving:** in each article retrieved from Wikipedia's database we must extract sentences relative to our input claim in order to apply a "kind of" comparison as well as present the winning sentences as evidence to the user.
- **Sentence Comparison:** here we must create a model or use a pre-existing natural language inferencing model in order to compare retrieved sentences from Wikipedia's articles and the input claim.
- **Credibility Evaluation:** not all informations in Wikipedia are true.

These can be regarded as main challenges of our Fact-Checking project, but, evidently, other problems can be presented for example the verifiability of claims, not all claims can be verifiable, especially if it is an personal opinion or a personal belief, in this case the model must not give a True or False assessment but it should tell the user that there is not enough information (NEI) to make such an assessment.

2.2 Related work and solutions

Chapter 3

Conclusion

Chapter 4

Perspectives

Bibliography

[1] Avery L. Wang. An industrial-strength audio search algorithm. In *Proceedings of the 4th Symposium Conference on Music Information Retrieval*, 2003.

[2] Peter Grosche, Meinard Müller, and Joan Serrà: Audio Content-Based Music Retrieval. In *Meinard Müller and Masataka Goto and Markus Schedl (ed.): Multimodal Music Processing, Schloss Dagstuhl—Leibniz-Zentrum für Informatik*, 2012.

[3] Audio Identification : https://www.audiolabs-erlangen.de/resources/MIR/FMP/C7/C7S1_AudioIdentification.html.

[4] J. Haitsma, T. Kalker, and J. Oostveen, "Robust Audio Hashing for Content Identification". In *n International Workshop on Content-Based Multimedia Indexing*, 2001.

[5] C.J. Burges, J. C. Patt, and S. Jana, "Distortion discriminant analysis for audio fingerprinting". In *IEEE Transaction on Speech and Audio Proc*, 2003.

[6] 6.050J/2.110J – Information, Entropy and Computation – Spring 2008 6.05https://mtlsites.mit.edu/Courses/6.050/2008/notes/mp3.html.

[7] Seeing circles, sines, and signals https://jackschaedler.github.io/circles-sines-signals/sound.html.

[8] Piotr Indyk, Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality.

[9] Jerome Schalkwijk, A Fingerprint for Audio https://medium.com/intrasonics/a-fingerprint-for-audio-3b337551a671.

[10] Jang et al. Pairwise Boosted Audio Fingerprint, 2009.

[11] Short-Time Fourier Transform. In *Sensor Technologies for Civil Infrastructures*, 2014.

[12] Nasser Kehtarnavaz. In *Digital Signal Processing System Design (Second Edition)*, 2008.