A proposal for the Machine Learning Application
Research Paper

March 22, 2020

Optimising online documents for fact-checking

Group:

Alpha

Members:

Alexander Peikert	alexpeikert@uni-koblenz.de	xxxxxxxxx
Clemens Steinmann	csteinmann@uni-koblenz.de	218200209
Erwin Letkemann	erwinletkemann@uni-koblenz.de	xxxxxxxxx
Julian Dillmann	juliandillmann@uni-koblenz.de	218100919

WeST
Universität Koblenz-Landau
Germany

1 Introduction

The aim of the application is to extract content of published news articles and to assign them various " Information nutritional labels", which are based on [1].
The web application can be used by different kind of users:

1. Ordinary web users can find similar articles for a given news article. The received labels can help answering the question whether the article was trustworthy or not. If latter, the user can continue reading on the topic with articles from different sources suggested by the application.

2. Other researchers can use our application, especially the labels, as a basis on further research on automation of fact-checking.

The labels we are focusing on will be at least 3 of the following: factuality / opinion, readability, emotion, controversy, credibility.

2 Application

2.1 Components

- Database
- Similarity checker / finder
- Content extractor
- Content NLP (Natural language processing)
- Web-application

2.2 General procedure

Given an input URL to an English news article, the applications' database will check whether the url has already been checked or not.

1. If the news article is present in our database we will output the already calculated labels and the x-best matching articles in our database.

2. Otherwise first the news article content extractor will crawl the html to extract the plaintext of the news article and second each of the labels will evaluated with corresponding natural language processing methods. The result will be stored added to the database witch leads us to 1.

2.3 Qualities of the components

- The **database** should hold: URL, metadata, labels, topic, published date, and should be quickly accessible based on topics to judge relevance of articles given another article.
- **Similarity** between two news articles describes the general topic in our use case. Consequently best matching news articles are those which have the highest lexical similarity and semantic similarity.
- **Labels** don't need to be specifically trained of/for news articles. Those are general purpose algorithms for each individual task.
- Possibly include the most used emotions to indicate discussion

on opposite viewpoints within an articles. (See basic idea of Emotion)

2.4 Basic ideas

most likely to change

- Factuality/ Opinion: [1] NLP to calculate ratio of factual to opinion sentences.
- Readability: [2] use some existing algorithm on the text.
- Emotion: [3], [4] NLP of sentences and sum will tell us the ratio of positive and negative emotions given in the text.
- Controversy: [5] checking for the number included Topics over the text.
- Credibility: [6] (or possibly some other existing website) compare to metadata.

3 Related Work

[1] Fuhr, Norbert et al. "**An Information Nutritional Label for Online Documents.**" SIGIR Forum 51 (2018): 46-66.

[2] Sahu, Ishan & Majumdar, Debapriyo. (2017). **Detecting Factual and Non-Factual Content in News Articles.** 1-12. 10.1145/3041823.3041837.

[3] **AFFIN Database** Informatics and Mathematical Modelling, Technical University of Denmark (2011)

[4] Cambria, Erik et al. "**The Hourglass of Emotions.**" COST 2102 Training School (2011).

[5] "Wikipedia:List of controversial issues" Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 18 March 2020, Web. 22 March 2020, en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues

[6] "Wikipedia:Reliable sources/Perennial sources" Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 21 March 2020, Web. 22 March 2020, en.wikipedia.org/wiki/Wikipedia:Reliable_sources/Perennial_sources