Optimising online documents for fact-checking

Group:

Alpha

Members:

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

A proposal for the Machine Learning Application
Research Paper

WeST
Universität Koblenz-Landau
Germany

March 22, 2020

1 Introduction

The aim of the application is to extract content of news articles and to assign labels based on [1] for each article. The web-based application will serve two purposes for two different kinds of users:
1. Ordinary web users can use the website to find similar articles for a given news article. The labels, with the help of self judgement, can be used as guideline for discovering new news sources or for evaluation of different news articles.
2. The second group are other researches. Our projects creates a basis for further research on the topic trustworthiness classification of news articles.
The labels we are focusing on will be at least 3 of the following: factuality/ opinion, readability, emotion, controversy, credibility.

2 Application

Component:
- Database
- Similarity checker/ finder
- Content extractor
- Content NLP
- Web-application

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 labels and find the 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.

Definitions??? / Other features
- The **database** should hold: the url, metadata, the labels, the topic. 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 the news articles with the highest sum of normalized labels on the same topic.
- **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)

Basic ideas likely is subject 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] NLP of sentences and sum will tell us the ratio of positive and negative emotions given in the text.
- Controversy: [4] checking for the number included Topics over the text.
- credibility: [5] (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] to many methods. (didn’t check)

[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