Because classification and labeling news articles based

Design

Because the issue of fake news detection on social media is both challenging and relevant, as a result of this work, literature and papers concerning natural language processing and social networks data mining were studied. This helped to achieve the goal of creating a service which may help to detect bias and falseness in social media publications and newspaper articles. This was achieved by reviewing existing literature in two phases: characterization and detection. In the characterization phase, we introduced the basic concepts and principles of fake news in both traditional media and social media. While studying the detection phase, we conducted a comprehensive review of existing fake news detection approaches from a data mining perspective, including feature extraction and construction of a model.

An overview of different natural language processing techniques was conducted, among which, models like Bag of Words, N-Gram and Binary co-occurrence coefficient were selected for the classification. As a result of conducted study, a list of requirements for a solution of the problem was created, namely, to create an automated software for classification and labeling news articles based on NLP feature extraction and to perform scoring of its performance and quality. Following up on that, a comprehensive solution for classifying and labeling various pieces of media was designed and programmed.

Various kinds of software testing of the developed system were performed which allowed to achieve a reduction of the number of potential inaccuracies and fall-throughs. A set of learning experiments was conducted in order to build precise classifiers. Created tool allows for a variety of deep analytical operations on the neuro part of natural language processing with the use of latest software tools available on the market. The created product is profiled for newspaper editorial offices, chief editors as well as solo journalists out there in the web who want to verify their news sources. It allows a user not only to input all the parts of an article inside a handy UI form, send it to server and obtain results in a nice and understandable manner but also allows a user of an API to process thousands of articles at once.

The application is completely modular and available for testing, and open source code will allow anyone to make the necessary suggestions and changes to the project. Implemented software has classification accuracy of 81% on the testing data set (while it is proven that humans are only 2% better than chance) which might be even more increased by further research and fine-tuning. While these numbers look promising for initial steps towards tackling the challenge that the problem of fake news possesses globally, one has to acknowledge that at least the quarter of the score may not be directly applied in a real world situation, due to the fact that the chosen data set was artificially created by randomly combining article bodies with article headline – for example, a headline which holds the text " Isis claims to behead US journalist" was combined with an article body which was about who is going to be the main actor in a biopic about Steve Jobs. Although this headline/article pair was (obviously) tagged as " unrelated", this is not something that is usually spread in a real-world scenario and is understood by a real fake news. For the more fine-grained classification of articles that have been classified as " related", the three-way classification is a relevant first step, but other classes may need to be added to the set, or a more detailed division may need to be made in order to proceed with tackling the fake news challenge. Additionally, we see the integration of known facts and general discourse knowledge (possibly through Linked Data), and the incorporation of source credibility information as important and promising suggestions for future research.