PanQura

User Experience Design and Credibility

3pc GmbH Neue Kommunikation, Condat AG, Semtation GmbH, DFKI GmbH, Fraunhofer FOKUS

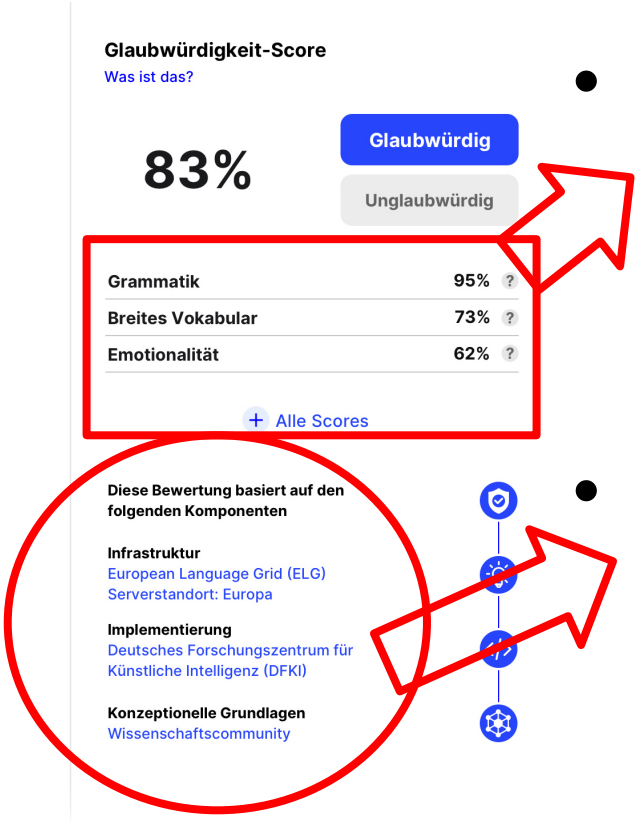
META-FORUM, 15.11.2021

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

qurator.ai/panqura/

User Experience of Credibility Ratings



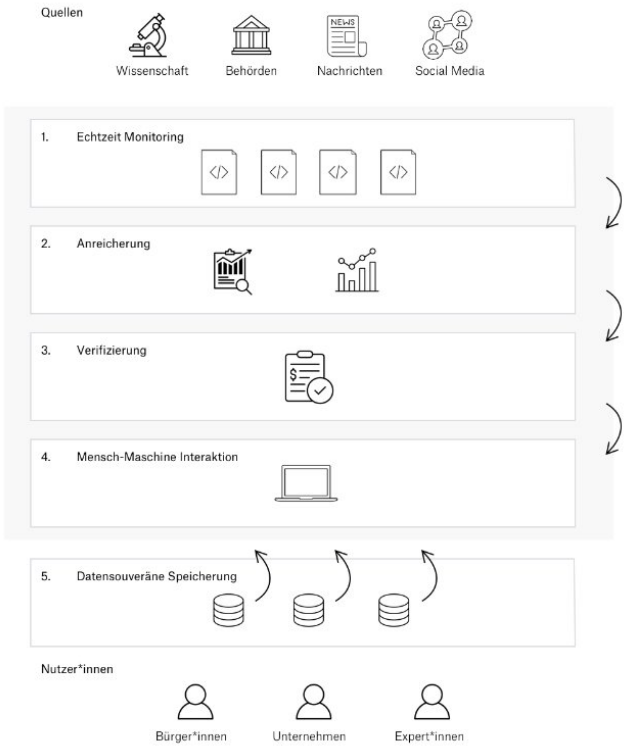
- **transparency** vs. **superabundance** of information
- **authority**: What is the source of that algorithm?

Main Findings

1. Some **metadata** (source of content) is more important, some less (source of algorithm).
2. **Authorship** is poorly defined for software as a service: concept, implementation, infrastructure, etc.
3. Automatic assessment can influence users‘ perception of credibility, but only if it is based on relevant **criteria** and backed by a respected **authority**.

Anhang

Meilenstein 1: Initialer Demonstrator



Ziel des Panqura-Projekts ist die Entwicklung einer Technologieplattform für mehr Informationstransparenz. Künftig stellt die Plattform eine Reihe von KI-basierten Werkzeugen zur vereinfachten Recherche Pandemie-bezogener Informationen bereit und unterstützt bei der Evaluation verfügbarer Internetquellen.

Mit dem Meilenstein 1 präsentiert das Bündnis einen ersten initialen Demonstrator. Er zeigt die anvisierten Funktionalitäten für die Erkennung und Evaluierung vom Themen, Fakten, Behauptungen und Glaubwürdigkeit auf und integriert sie in eine Reihe von Use Cases.

Content-Focused Webpage Credibility Evaluation Using W3C Credibility Signals

Goal: Development of an application exposed through Rest API to assess the credibility of webpages by evaluating a range of credibility signals - webpage properties used as credibility indicators

Extract relevant data from webpage

Compute credibility signal sub-scores

Combine sub-scores into webpage credibility score

Credibility Signals

- Analyse headline, text body, links and some HTML content (e. g., whether there are authors specified)
- Focus on signals intrinsic to content, such that the same content would be evaluated equally on different websites, and adversarial measures are harder
- Many signals related to readability and language structure (readability grades, word counts, average word lengths…)
- Additionally, among others:
  - Headline clickbait classification
  - Grammar/spelling errors
  - Emotionality & subjectivity
  - Vocabulary
  - Punctuation & use of all-caps

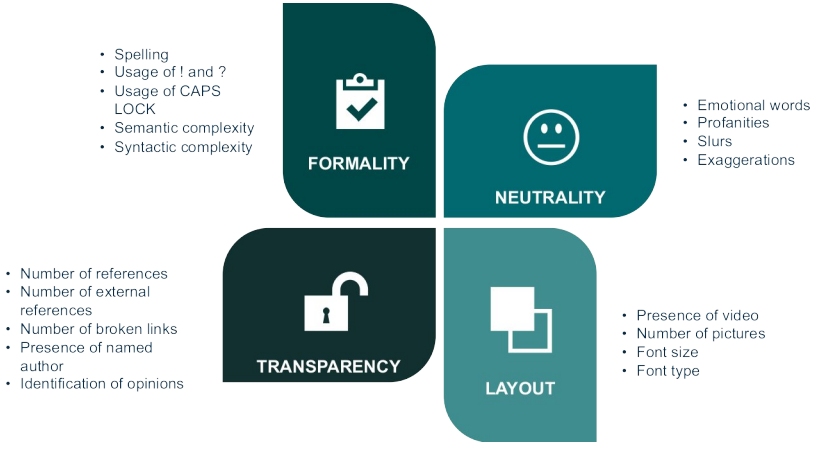
Preliminary Results

- Weights for combination of signal sub-scores into final webpage score are derived from previous scientific findings and own analysis of signal statistics on data sets
- Conflicting scientific results on correlation with credibility for some signals (e. g., question mark usage in text)
- Some signals that are mentioned in the literature are (almost) irrelevant due to non-occurrence
  (e. g. profanity, grammar/spelling errors)
- Some well-performing signals are not included in the W3C WebCred credibility signal list, likely due to being very specific and/or difficult to gauge intuitively
  (e. g., type-token-ratio, average word length)

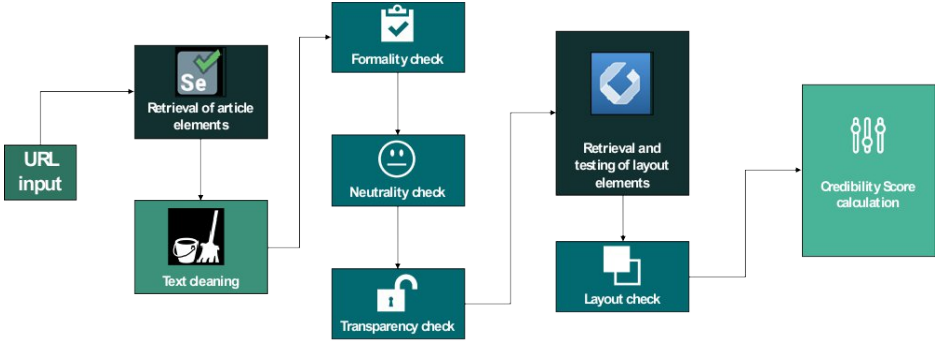
Master-Arbeit
Léon Avilés

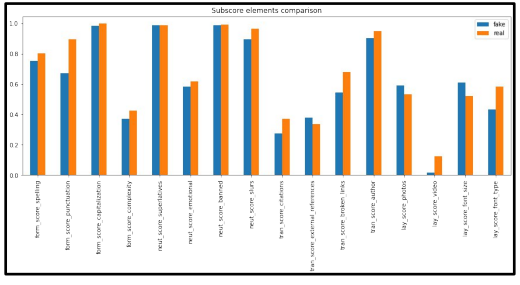
Credibility Score using W3C Signals and Metrics

Signals and Metrics



Program Architecture

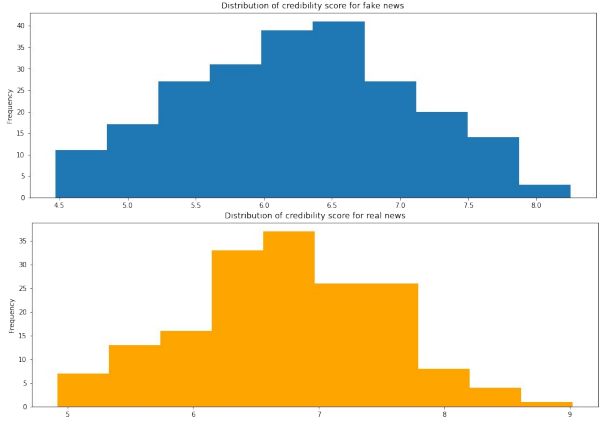




Results

Best performance:

- Usage of ! and ? ☾ 22.4%
- Font type (serif fonts) ☾ 14.9%
- Presence of references ☾ 13.4%
- Presence of video ☾ 10.5%
- Broken links ☾ 9.6%



- Distribution shows shift towards higher scores for real news
- Example: 37.8% of fake news and only 17.5% of real news have scores below 6

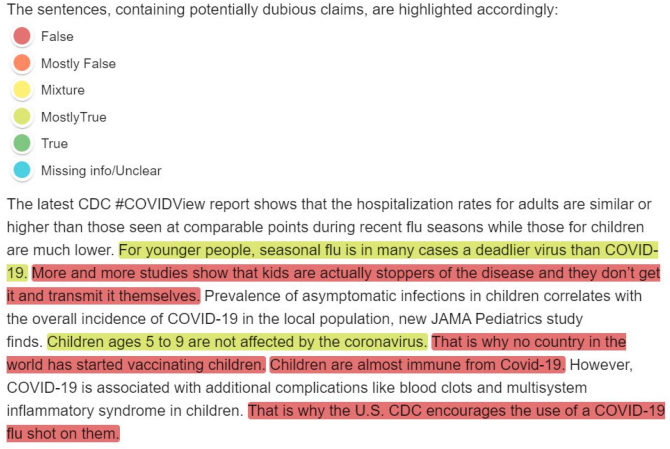
Master-Arbeit
Eliza Danila

Fact Checking Using Trusted Knowledge Bases

- Goal: a high-performance component for fact checking of small- to medium-sized documents on the topic of COVID-19
- Component pipeline:
  - Parse text document into sentences
  - Fake news detection (classification in *suspicious* and *regular sentences* using Transformer models, fine-tuned on a custom dataset)
  - Claim extraction from the suspicious sentences (via spaCy NLP library)
  - Claim verification (via Google Fact Check Tools API)
  - Mapping textual to a numerical rating of each claim
  - Visualizations: Streamlit app with a custom Vue + Vuetify frontend component
- Overall accuracy of 98.1% achieved in the sequence

Classification task using DistilBERT, compared to

95.1% with a simple LSTM implementation



Bachelor-Arbeit
Elitsa Pankovska

Political Bias Classification

- Using combinations of features (BOW, TF-IDF and BERT) and models (LR, NB, RF and EasyEnsemble), we get the best results with a Random Forest classifier using BERT representations of the input.
- Per class performance illustrates that both extremes (far-left, far-right) are the easiest to classify despite low number of support cases.
- Approach performs comparable to the top-5 of the 2019 Hyperpartisan News Detection task, with 0.67 F1 (vs. 0.43 with multi-class setup) on this data set.

Model	BOW	TF-IDF	BERT
Logistic Regression	0.3132	0.2621	0.3389
Naive Bayes	**0.4243**	0.2234	0.3637
Random Forest	0.4007	**0.4303**	**0.3836**
EasyEnsemble	0.4197	0.4070	0.3432

Class	Precision	Recall	F_1	Support
Far-left	0.59	0.40	0.48	215
Centre-left	0.34	0.38	0.36	1,159
Centre	0.31	0.23	0.27	1,349
Centre-right	0.51	0.55	0.53	1,754
Far-right	0.46	0.58	0.51	671
Total	**0.44**	**0.43**	**0.43**	**5,148**

- Demonstrates the increased difficulty when using multi-class labels (5-point scale).
- If quality and transparency are important, more fine-grained classification is necessary.
- *Accepted for publication at WOAH 2021 (Workshop on Online Abuse and Harms 2021)*

Assessing COVID-related News

- **Sources (selection)**
  - Credibility Signals (W3C Credible Web Community Group)
  - Fact Check APIs (for example, Google‘s)
  - Political bias classification
  - *Additional classifiers*
- **Mixed Method Approach**
  - Deep Learning
  - Linguistic & formal heuristics
  - External knowledge bases
- **Infrastructure**
  - QURATOR- and ELG-compatible Language Technology service
  - Easy access through a simple user interface
  - Cross-platform, with cloud capabilities

Idea: combine these into an ensemble of services

deploy through QURATOR/ELG platforms for PANQURA prototype

Qurator
Curation Technologies

EUROPEAN LANGUAGE GRID

Evaluating pre-trained, domain-specific vs. general Transformer models on expert and non-expert questions about COVID-19

Research goal: The objective of this project is to evaluate whether a prior distinction between expert and non-expert questions about COVID-19 before choosing a question-answering model can increase the quality of the predicted answers

Data

Questions: EPIC-QA question sets (43 non-expert, 45 expert questions)
Answers: COVID-19 Open Research Dataset (CORD-19)[2] (subset of 100 non-expert and 100 expert documents used in current project)

Methods

Models

- *BERT Base* (Devlin et al. (2019)): trained on general-domain
- *BioBERT* (Lee et al. (2020): trained on medical articles

Further filtering

- Problem: too long, answer quality
- Idea: apply additional filters
- Applied filters: limitation on answer length (50 tokens), Levenshtein distance, vector similarity between Q and A

Experimental setup

- 8 setups in total: 2 models and 4 setups with filters per question set

Question / Answer analysis

- Questions analyzed w.r.t. three aspects (Pomerantz, J. (2005)): "wh"-words, subject / vocabulary, function of expected / correct answer
- Qualitative evaluation of Answers (no gold answers available): answer score (Oniani, Wang (2020): 5 scores of answer quality (*5-relevant*, *4-well-formed*, *3-informative*, *2-acceptable*, *1-poor*)
- Further analysis w.r.t. Levenshtein distance and embedding similarity between question and answer, answer length and function

Preliminary Results

- Best performing models (w.r.t answer score): BioBert with Levenshtein distance filter and BioBert with vector similarity filter for expert questions (average score: 3.88)
- Poor quality across both question sets, perhaps due to small data set size and noisy data
- In general, answer quality is higher for expert questions

Studienarbeit
Karina Hensel