Fuhr, Norbert et al.

Article

An Information Nutritional Label for Online Documents

ACM SIGIR Forum

Suggested Citation: Fuhr, Norbert et al. (2017) : An Information Nutritional Label for Online Documents, ACM SIGIR Forum, ISSN 0163-5840, ACM, New York, NY, Vol. 51, Iss. 3, pp. 46-66

ARTICLE

An Information Nutritional Label for Online Documents

Norbert Fuhr,[1] University of Duisburg-Essen
Anastasia Giachanou,[2] Università della Svizzera italiana (USI)
Gregory Grefenstette,[3] Institute Human Machine Cognition
Iryna Gurevych,[4] Technische Universität Darmstadt
Andreas Hanselowski,[5] Technische Universität Darmstadt
Kalervo Jarvelin,[6] University of Tampere
Rosie Jones,[7] Consultant
Yiqun Liu,[8] Tsinghua University
Josiane Mothe,[9] Université de Toulouse, (IRIT)
Wolfgang Nejdl,[10] Leibniz Universität Hannover,
Isabella Peters,[11] ZBW Leibniz Information Centre for Economics and Kiel University
Benno Stein,[12] Bauhaus-Universität Weimar

Abstract

With the proliferation of online information sources, it has become more and more difficult to judge the trustworthiness of news found on the Web. The beauty of the web is its openness, but this openness has lead to a proliferation of false and unreliable information, whose presentation makes it difficult to detect. It may be impossible to detect what is "real news" and what is "fake news" since this discussion ultimately leads to a deep philosophical discussion of what is true and what is false. However, recent advances in natural language processing allow us to analyze information objectively according to certain objective criteria (for example, the number of spelling errors). Here we propose creating an "information nutrition label" that we can automatically generated for any online text. Among others, the label provides information on the following computable criteria: factuality, virality, opinion, controversy, authority, technicality, and topicality. With this label, we hope to help readers make more informed judgments about the items they read.

[1] e-mail:norbert.fuhr@uni-due.de
[2] e-mail:anastasia.giachanou@usi.ch
[3] e-mail:ggrefenstette@ihmc.us
[4] e-mail:gurevych@ukp.informatik.tu-darmstadt.de
[5] e-mail:hanselowski@aiphes.tu-darmstadt.de
[6] e-mail:kalervo.jarvelin@uta.fi
[7] email:rosie.jones@acm.org
[8] e-mail:yiqunliu@tsinghua.edu.cn
[9] e-mail:josiane.mothe@irit.fr
[10] e-mail:nejdl@kbs.uni-hannover.de
[11] e-mail:i.peters@zbw.eu
[12] e-mail:benno.stein@uni-weimar.de

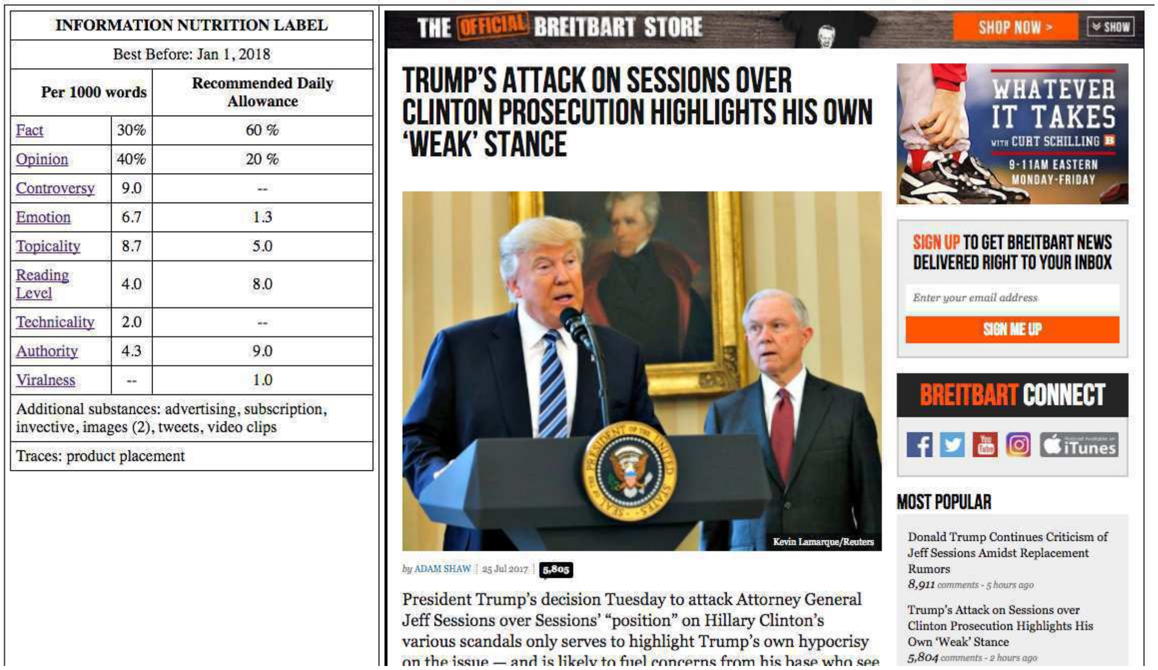

Figure 1: Mockup of the envisaged information nutrition label.

1 Introduction

The 2016 American presidential elections were a source of growing public awareness of what has been termed "fake news," a term is used to describe the observation that "in social media, a certain kind of 'news' spread much more successfully than others, and, that these 'news' stories are typically extremely one-sided (hyperpartisan), inflammatory, emotional, and often riddled with untruths" [20].

Claims in news can take various forms. In the form of a verifiable assertion ("The density of ice is larger than the density of water.") we have a fact checking situation, which can be clarified given access to online dictionaries or encyclopedias. In the form of a non-verifiable or not easily verifiable assertion ("Hillary Clinton is running a child sex ring out of a D.C.-area pizza restaurant.", "Marijuana is safer than alcohol or tobacco.") one has to take a stance, i.e., the reader has to decide whether she believes the claim or not. Such a decision can neither universally nor uniquely be answered by means of a knowledge base but is to be clarified on an individual basis and may undergo change over time.

To help the online information consumer, we propose an Information Nutrition Label, resembling nutrition fact labels on food packages. Such a label describes, along a range of agreed-upon dimensions, the contents of the product (an information object, in our case) in order to help the consumer (reader) in deciding about the consumption of the object. The observations above however also imply a particular, self-imposed, ethical limitation of our concept:

> *(Our manifest) It is not our intention to say what is true or what is false, right or wrong, and in particular not what is good or bad. That is, an Information Nutrition Label is not a substitute for a moral compass.*

Thus, as technical consequence, we do not propose a system that would state that a piece of news is true or false, leaving that decision up to the final user. Ultimately, just as with a food label, it is up to the consumer to consult the information nutrition label and to decide whether to consume the information or not.

In addition to aiding a consumer's decision making process, we also see possible technical uses as well as societal impacts of our Information Nutrition Label:

- personalized relevance ranking for search engine results
- information filtering according to personal preferences
- machine-based fake news detection
- learning and teaching of information assessment
- raising awareness and responsibility about deciding what to read.

2 An Information Nutrition Label

Of course, the assessment of information is not a new discipline. There is a large body of research related to the concept of "information quality", for which Levis et al. provide a useful overview [15]. While there is no unique definition for the concept, information quality is usually interpreted in terms of utility, or the "fitness for use in a practical application" [27]. Note that our paper will neither reinterpret nor extend this concept of quality; instead, we are aiming at a practical means to ease information consumption and meta reasoning when given an online document by breaking down a quality judgment into smaller, measurable components.

We consider the Wikipedia quality endeavour as the most related precursor to our proposal. Aside from its rather informal quality guidelines, Wikipedia has formalized its quality ideal with the so-called featured article criteria[1], and, even more important, distinguishes more than 400 quality flaws to spot article deficits [2]. In particular, the machine-based analysis of Wikipedia articles to detect flaws in order to assess article quality [3] corresponds closely to our idea of computing the separate dimensions of an information nutrition label. However, because of our use case, the nutrition label dimensions as well as their computation differs from the Wikipedia setting.

The following sections describe measurable qualities that may be included in such an information nutrition label and that we consider valuable in order to assess the nutrient content of an information object. Each of these categories have been the subject of more or less detailed experimentation in the natural language processing, information retrieval, or web sciences communities:

- factuality
- readability
- virality
- emotion
- opinion
- controversy
- authority / credibility / trust
- technicality
- topicality

In the next sections we will describe each category, and the natural language processing tasks for measuring them. Some tasks have been extensively studied and some are recent suggestions. For each category, we will give a brief description of the task; current methods for producing the measure when

[1]Wikipedia, "Featured articles," last modified February 19, 2017, `http://en.wikipedia.org/wiki/Wikipedia:Featured_articles`

they exist, or suggestions of how they could be measured otherwise; existing data sets that have either been developed for this task, or that can be used in developing new methods for the task; each section ends with a list of further reading. Some tasks have been extensively studied and some are recent suggestions.

3 Factuality

3.1 Task for Factuality Assessment

Factual statements are different from expressions of opinion. An author who claims to be presenting facts writes in a certain way that can be detected.

The task of determining the level of commitment towards a predicate in a sentence according to a specific source, like the author, is typically addressed as factuality prediction [22]. Lexical cues, such as modals *will*, *shall*, *can* indicate the confidence of the source whether a proposition is factual. However, in contrast to a binary decision, the underlying linguistic system forms a continuous spectrum ranging from factual to counterfactual. Thus, for the assessment of the factuality for the whole document, one needs to compute the average factuality of all the propositions contained in the text.

Since we are not planning, in this category of our Information Nutrition Label, to judge the truthfulness of the statements in a given text, as it is attempted in the domain of automated fact checking, we are only interested in determining whether a statement is factual from the perspective of the author. The issue of whether the statements in the documents are controversial and may therefore not be reliable, is discussed in section 8 about Controversy.

3.2 Methods for Factuality Assessment

For factuality prediction rule-based approaches as well as methods based on machine learning have been developed.

The De Facto factuality profiler [23] and the TruthTeller algorithm [17] are rule-based approaches, which assign discrete scores of factuality to propositions. In the process, dependency parse trees are analyzed top-down and the factuality score is altered whenever factuality affecting predicates or modality and negation cues are encountered.

A machine learning based approach has been applied to factuality prediction in [14]. The authors used a support vector machine regression model to predict continuous factuality values from shallow lexical and syntactic features such as lemmas, part-of-speech tags, and dependency paths.

The rule-based approach has been combined with the machine learning based method in [25]. Thereby, the outputs from TruthTeller were used as linguistically-informed features for a support vector machine regression model in order to predict the final factuality value.

3.3 Data sets for Factuality Assessment

There are a number of annotation frameworks, which have been suggested to capture the factuality of statements. On the basis of the suggested annotation schemes, a number of data sets have been constructed.

Fact-Bank [22] is a corpus which was annotated discretely by experts according to different classes of factuality: Factual, Probable, Possible, Unknown. In this corpus, factuality has been assessed with respect to the perspective of the author or discourse-internal sources.

The MEANTIME corpus was introduced in [18] and was also annotated discretely by expert annotators. The propositions have been classified as Fact / Counterfact, Possibility (uncertain), Possibility (future) with respect to the author's perspective.

The UW corpus [14] was annotated on the basis of a continuous scale ranging from -3 to 3. The annotation was performed by crowd workers who judged the factuality score from the author's perspective.

In [25], the annotation schemes of the three different corpora have been merged in order to combine the three data sets into one single large corpus. For this purpose, the discrete scales used for the Fact-Bank and MEANTIME corpora have been mapped to the continuous scale of the UW corpus.

3.4 Further reading for Factuality Assessment

1. Nissim Malvina, Paola Pietrandrea, Andrea Sanso, and Caterina Mauri. "Cross-linguistic annotation of modality: a data-driven hierarchical model." In Proceedings of the 9th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation, pp. 7-14. 2013.

2. O'Gorman Tim, Kristin Wright-Bettner, and Martha Palmer. "Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation." In Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016). 2016.

3. Ghia Elisa, Lennart Kloppenburg, Malvina Nissim, Paola Pietrandrea, and Valerio Cervoni. "A construction-centered approach to the annotation of modality." In Proceedings of the 12th ISO Workshop on Interoperable Semantic Annotation, pp. 67-74. 2016.

4. Guggilla Chinnappa, Tristan Miller, and Iryna Gurevych. "CNN-and LSTM-based Claim Classification in Online User Comments." In Proceedings of the COLING 2016.

5. Szarvas Gyrgy, Veronika Vincze, Richrd Farkas, Gyrgy Mra, and Iryna Gurevych. "Cross-genre and cross-domain detection of semantic uncertainty." Computational Linguistics 38, no. 2 (2012): 335-367.

4 Readability

4.1 Task for Readability Measurement

Readability is defined in Wikipedia as "the ease with which a reader can understand a written text." We would like the Information Nutrition Label to provide the potential reader some idea about this aspect of the text.

Readability can be measured by the accuracy of reading and the reading speed for the reader. Readability depends mainly on three categories of factors: writing quality, targeted audience and presentation.

Writing quality refers to the grammatical correctness of the text (morphology, syntax) such as taught in elementary schools [28]. Readability also depends on the *target audience* or in other words the level of educational background the reader needs to have to understand the text content (the complexity of its vocabulary and syntax, the rhetorical structure). Finally, the *presentation* refers to typographic aspects like font size, line height, and line length [5] or visual aspects like color [13].

4.2 Methods for Readability Measurement

Collins-Thompson provides a recent state of the art summary of automatic text readability assessment [7]. Two main factors are used in readability measures: the familiarity of semantic units (vocabulary) and the complexity of syntax.

Automatic readability measures estimate the years of education or reading level required to read a given body of text using surface characteristics. The current measures are basically linear regressions based on the number of words, syllables, and sentences [19] [7].

Wikipedia presents a number of readability tests in their eponymous article that usually involve counting syllables, word length, sentence length, and number of words and sentences.[2]

Crossley et al. developed Coh-Metrix, a computational tool that measures cohesion and text difficulty at various levels of language, discourse, and conceptual analysis [8]. De Clercq et al. proposed to use the crowd to predict text readibility [10].

Automatic methods have been developed for different languages as for Arabic [1], French [12], Polish [6], or Spanish [24] to cite a few.

4.3 Data sets for Readability Measurement

There are a number of data sets sets and sample demos for readability measurement. The data sets include:

- Text Exemplars and Sample Performance Tasks in Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects (183 pages). [Copyright and Permissions] Includes examples with different grades, genres (English only).[3]
- [11] mentions a collection of Weekly Reader extracts that may still be available.
- Math Webpage Corpus with Readability Judgments[4]

Sample demos:

- Readability[5] implemented by Andreas van Cranenburgh (`andreasvc` on github) calculates a number of standard reading level features, including Flesch, Kincaid and Smog (a descendent of an nltk_contrib package [6]). This package expects sentence-segmented and tokenized text. For English, van Cranenburgh recommends tokenizer.[7] For Dutch, he recommends the tokenizer that is part of the Alpino parser[8]. There is also *ucto*[9], a general multilingual tokenizer. One can also use the tokenizer included in the Stanford NLP package.

[2] https://en.wikipedia.org/wiki/Readability#Popular_readability_formulas, last access Oct. 11, 2017
[3] http://www.corestandards.org/assets/Appendix_B.pdf
[4] https://web.archive.org/web/*/http://wing.comp.nus.edu.sg/downloads/mwc
[5] https://pypi.python.org/pypi/readability
[6] https://github.com/nltk/nltk_contrib/tree/master/nltk_contrib/readability
[7] http://moin.delph-in.net/WeSearch/DocumentParsing
[8] http://www.let.rug.nl/vannoord/alp/Alpino/
[9] http://ilk.uvt.nl/ucto

Test cases:

```
$ ucto -L en -n -s ''
"CONRAD, Joseph - Lord Jim.txt" |
readability
[...]
readability grades:
Kincaid:                  4.95
ARI:                      5.78
Coleman-Liau:             6.87
FleschReadingEase:       86.18
GunningFogIndex:          9.4
LIX:                     30.97
SMOGIndex:                9.2
RIX:                      2.39
```

Other tools: Benchmark Assessor Live[10], and also see Further Reading, below.

4.4 Further reading for Readability Measuring

1. Flesch and Kincaid Readability tests, [11] and the Wikipedia article on Readability[12] (for several other readability formulas)

2. Heilman, Michael, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. "Combining lexical and grammatical features to improve readability measures for first and second language texts." In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pp. 460-467. 2007.

3. Collins-Thompson, Kevyn. "Computational assessment of text readability: A survey of current and future research." ITL-International Journal of Applied Linguistics 165, no. 2 (2014): 97-135.

4. De La CHICA, Sebastian, Kevyn B. Collins-Thompson, Paul N. Bennett, David Alexander Sontag, and Ryen W. White. "Using reading levels in responding to requests." U.S. Patent 9,600,585, issued March 21, 2017.

5. Vajjala, Sowmya, and Detmar Meurers. "On the applicability of readability models to web texts." In Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations, pp. 59-68. 2013.

6. Rello, Luz, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. "Frequent words improve readability and short words improve understandability for people with dyslexia." In IFIP Conference on Human-Computer Interaction, pp. 203-219. Springer, Berlin, Heidelberg, 2013.

[10] https://www.readnaturally.com/assessment-tools
[11] https://en.wikipedia.org/wiki/Flesch%E2%80%93Kincaid_readability_tests
[12] https://en.wikipedia.org/wiki/Readability

(excerpt: ... To determine how much individual queries differ in terms of the readability of the documents they retrieve, we also looked at the results for each query separately. Figure 4 shows the mean reading level of the Top-100 results for each of the 50 search queries...)

7. Newbold, Neil, Harry McLaughlin, and Lee Gillam. ”Rank by readability: Document weighting for information retrieval.” Advances in multidisciplinary retrieval (2010): 20-30. (”...Web pages can be, increasingly, badly written with unfamiliar words, poor use of syntax, ambiguous phrases and so on....”)

8. Feng, Lijun, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. ”A comparison of features for automatic readability assessment.” In Proceedings of the 23rd International Conference on Linguistics: Posters, pp. 276-284. Association for Computational Linguistics, 2010.

5 Virality

Sometimes fake news “catches on” in a viral way, spreading from one dubious source across the internet into the mainstream. We are interested in including some measure of virality in our Information Label.

In analyses of information objects and information flows on the internet the notion of “virality” is often stressed. This is especially true when discussed in the particular context of marketing and advertisement. Virality means that “information objects spread the way that viruses propagate. [Hence,v]irality has become a common way to describe how thoughts or information move through a human population, or the internet, social network sites in particular” [13]. The metaphor of the virus supports consideration of different properties that may influence the spread of information but that can also be used to quantify virality.

5.1 Task for Virality Detection

For the detection of virality in texts or other information objects four types of property sets have to be taken into account: a) the sender, b) the information object, c) the recipient, and d) the channel. The combination of these sets influences the speed with which a virus spreads and also determines how far it can reach. The major factors on the sender side are their popularity and authority, the size of their network, but also the amount of trust they receive from recipients. The recipient must be able to receive the information object and should not be immune to it, e.g. because they had the information object before. The information object itself is often admissable to many different types of recipients, for example, because of its short topical distance to knowledge the recipients already hold. The channel offers varying functionalities and allows for different ease of use to further spread the information object. Higher ease of use encourages the sharing of information objects, e.g., retweeting a tweet on Twitter. Moreover, the environment in which the information object spreads is of interest, too. It may have been influenced by a frame setting activity, i.e. bringing certain information to the awareness of many recipients, that increases the probability of recipients getting infected, e.g. because they search for this type of information. The virality of information objects could also be subject to within-platform as well as cross-platform properties.

[13] https://en.wikipedia.org/wiki/Viral_phenomenon

5.2 Methods for Virality Detection

The determination of virality needs to operationalize all of these factors, especially with regard to the graph-like structure of the information flow. In social media, many signals can be used for this, e.g., number of likes, retweets, and comments, characteristics of followers, communities, or hashtags, or time of posting. Those factors build the ground for virality measurement. However, it is not only the quantity of these signals that may determine virality but also the speed with which information objects spread and how far they reach (e.g., when different communities are infected by the same information object).

5.3 Tools and Data for Virality Detection

Examples for existing software that visualizes the spread of claims (i.e. Hoaxy) or that follows memes are provided by the Indiana University Network Science Institute (IUNI) and the Center for Complex Networks and Systems Research (CNetS)[14].

There are also several data sets available that can be used for training, for example viral images[15] or tweets[16], see also Weng *et al.* in the Further Reading section.

5.4 Further reading for Virality

1. Weng, Lilian, Filippo Menczer, and Yong-Yeol Ahn. ”Virality prediction and community structure in social networks.” Scientific reports 3 (2013): 2522.

2. Weng, Lilian, and Filippo Menczer. ”Topicality and impact in social media: diverse messages, focused messengers.” PloS one 10, no. 2 (2015): e0118410.

3. Guerini, Marco, Carlo Strapparava, and Gözde Özbal. ”Exploring Text Virality in Social Networks.” In ICWSM. 2011.

4. Guille, Adrien and Hacid, Hakim and Favre, Cecile and Zighed, Djamel A. ”Information diffusion in online social networks: A survey.” ACM Sigmod Record 42.2 (2013): 17-28.

6 Emotion

6.1 Task for Emotion Detection

One characteristic of Fake News is that it may make an inflammatory emotional appeal to the reader. Emotional arguments often employ words that are charged with positive or negative connotations (such as *bold* or *cowardly*). Such language also appears in product and movie reviews.

The task here is to detect the sentences which are emotive in a document, and to calculate the intensity, the polarity and the classes of the affect words found there. The emotional impact of a document can either be averaged over the number of words, or be calculated by using some maximum value encountered [4].

[14]http://truthy.indiana.edu
[15]https://github.com/ArturoDeza/virality
[16]http://carl.cs.indiana.edu/data/#virality2013

6.2 Methods for Emotion Detection

As a sample method, an emotion detection method can include the following steps:

1. Divide document into sentences
2. Extract words, terms, negations, intensifiers, emoticons, parts of speech, punctuation from the sentence
3. Use these extracted items as features to classify the sentence
4. Identify which sentences carry emotion, and what emotion
5. Combine measures from all sentences to create a single emotion rating of the document.

6.3 Data sets for Emotion Detection

Data resources for emotion detection include sentiment lexicons and test/training data sets. Some of the former are:

- A list of Affect Lexicons[17] maintained by Saif Mohammad
- SenticNet[18]
- AFINN[19]
- List of affect resources[20] maintained by Bing Liu
- Affective Norms for English Words (ANEW) is a set of normative emotional ratings for 2,476 English words. We use the valence rating considering positive (respectively, negative) the ratings above (respectively, below) the mean.
- General Inquirer is a list of 1,915 words classified as positive, and 2,291 words classified as negative.
- MicroWNOp is a list of 1,105 WordNet synsets (cognitive synonyms) classified as positive, negative, or neutral.
- SentiWordNet assigns to each synset of WordNet (around 117,000) a positive and negative score determined by a diffusion process.
- Bias Lexicon is a list of 654 bias-related lemmas extracted from the edit history of Wikipedia [21]. Sentiment words are used as contributing features in the construction of this bias lexicon.

[17]http://saifmohammad.com/WebPages/lexicons.html
[18]http://sentic.net/downloads/
[19]http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010
[20]https://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html

Test and training data sets include: Reviews;[21] Twitter in 15 languages;[22] Twitter and emotions;[23] Twitter tweets;[24] Blog sentences;[25] Facebook statuses, CNN, the New York Times, Guardian, BBC news, ABC news;[26] three emotional dimensions (Valence, Arousal and Dominance)[27]

6.4 Further reading for Emotion Detection

1. Valitutti, Alessandro, and Carlo Strapparava. ”Interfacing WordNet-affect with OCC model of emotions.” In The Workshop Programme, p. 16. 2010.[28]

2. Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. ”Sentiment analysis algorithms and applications: A survey.” Ain Shams Engineering Journal 5.4 (2014): 1093-1113.

3. Giachanou, Anastasia, and Fabio Crestani. ”Like it or not: A survey of twitter sentiment analysis methods.” ACM Computing Surveys (CSUR) 49, no. 2 (2016): 28.

4. Cambria, Erik. ”Affective computing and sentiment analysis.” IEEE Intelligent Systems 31, no. 2 (2016): 102-107.

5. Tripathi, Vaibhav, Aditya Joshi, and Pushpak Bhattacharyya. ”Emotion Analysis from Text: A Survey.”[29]

7 Opinion

Opinion is an element of the text which reflects the author’s opinion, and readers’ opinions may differ. The output is a percentage, based on the fraction of words or sentences which are opinion, in contrast to facts. Authors of opinionated text may be surreptitiously pushing a certain viewpoint which is not explicitly expressed in the text.

7.1 Task for Opinion Detection

For the Information Nutrition Label, our task is to detect sentences that are opinionated, and calculate the percentage of opinionated sentences for entire text. Table 1 gives some examples of opinionated and factual sentences.

[21] https://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html#datasets
[22] https://www.clarin.si/repository/xmlui/handle/11356/1054
[23] http://saifmohammad.com/WebPages/EmotionIntensity-SharedTask.html
[24] http://www.sananalytics.com/lab/twitter-sentiment/
[25] https://inclass.kaggle.com/c/si650winter11/data
[26] https://github.com/minimaxir/interactive-facebook-reactions/tree/master/data
[27] https://github.com/JULIELab/EmoBank/tree/master/corpus
[28] https://source.opennews.org/articles/analysis-emotional-language/
[29] http://www.cfilt.iitb.ac.in/resources/surveys/emotion-analysis-survey-2016-vaibhav.pdf

Sentence	Label
The first amendment includes the most misused freedom in our country, which is the freedom of the press.	`Opinionated`
The 18th amendment to the constitution prohibited the manufacture, sale, or transportation of alcohol.	`Fact`
The 16th amendment gave congress to collect taxes from American citizens, and they have been collecting way too many taxes ever since	`Opinionated`
Result	Opinion-Ratio = 2/3

Table 1: Examples of Fact vs Opinion sentences as taught to US Elementary School Children[31], along with a score which could be computed from them.

7.2 Existing Methods for Opinion Detection

There is software available for opinion detection. Here are some:

- OpeNER[32] "aims to be able to detect and disambiguate entity mentions and perform sentiment analysis and opinion detection on the texts[33]..."
- Opinion Finder[34], see Wilson et al, in Further Readings below.
- Opinion Sentence Finder[35]. See also Rajkumar et al., below.
- NLTK opinion lexicon reader[36].

7.3 Data sets for Opinion Detection

There are also data sets for opinion detection:

- Fact vs. opinion as taught to US Elementary School Children.[37] These examples have answers[38], too. The overall output score is the percent of sentences which contain opinions.
- Bitterlemon collection 594 editorials about the Israel-Palestine conflict, 312 articles from Israeli authors and 282 articles from Palestinian authors.
- Opinion lexicon[39]

[32] http://www.opener-project.eu/
[33] http://www.opener-project.eu/getting-started/#opinion-detection
[34] http://mpqa.cs.pitt.edu/opinionfinder/opinionfinder_2/
[35] http://cse.iitkgp.ac.in/resgrp/cnerg/temp2/final.php
[36] http://www.nltk.org/_modules/nltk/corpus/reader/opinion_lexicon.html
[37] http://www.shsu.edu/txcae/Powerpoints/prepostest/fact1postest.html
[38] http://www.shsu.edu/txcae/Powerpoints/prepostest/fact1postans.html
[39] https://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html#lexicon

- Multi perspective question answering lexicon[40] corpus contains news articles and other text documents manually annotated for opinions and other private states (i.e., beliefs, emotions, sentiments, speculations, etc.).
- Arguing Lexicon[41]: includes patterns that represent arguing.

7.4 Further reading for Opinion Detection

1. Fact vs opinion as taught to US Elementary School Children[42]
2. Paul, Michael J., ChengXiang Zhai, and Roxana Girju. "Summarizing contrastive viewpoints in opinionated text." In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 66-76. Association for Computational Linguistics, 2010.
3. Yu, Hong, and Vasileios Hatzivassiloglou. "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences." In Proceedings of the 2003 conference on Empirical methods in natural language processing, pp. 129-136. Association for Computational Linguistics, 2003. "classify sentences as fact / opinion using word n-grams, word polarity"
4. Liu, Bing, Minqing Hu, and Junsheng Cheng. "Opinion observer: analyzing and comparing opinions on the web." In Proceedings of the 14th international conference on World Wide Web, pp. 342-351. ACM, 2005.
5. Wilson, Theresa, David R. Pierce, and Janyce Wiebe. "Identifying opinionated sentences." In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations-Volume 4, pp. 33-34. Association for Computational Linguistics, 2003.
6. Rajkumar, Pujari, Swara Desai, Niloy Ganguly, and Pawan Goyal. "A Novel Two-stage Framework for Extracting Opinionated Sentences from News Articles." In TextGraphs@ EMNLP, pp. 25-33. 2014.

8 Controversy

Controversy is a state of prolonged public dispute or debate, usually concerning a matter of conflicting opinion or point of view. The word was coined from the Latin controversia, as a composite of controversus "turned in an opposite direction," from contra "against" and vertere to turn, or versus (see verse), hence, "to turn against." The most applicable or well known controversial subjects, topics or areas are politics, religion, philosophy, parenting and sex (see Wikipedia articles in Further Reading, as well as Aharoni et al.) History is similarly controversial. Other prominent areas of controversy are economics,

[40] mpqa.cs.pitt.edu/corpora/mpqa_corpus/
[41] http://mpqa.cs.pitt.edu/lexicons/arg_lexicon
[42] http://teaching.monster.com/training/articles/2589-k-5-fact-versus-opinion

science, finances, culture, education, the military, society, celebrities, organisation, the media, age, gender, and race. Controversy in matters of theology has traditionally been particularly heated, giving rise to the phrase *odium theologicum*. Controversial issues are held as potentially divisive in a given society, because they can lead to tension and ill will, and as a result they are often considered taboo to be discussed in the light of company in many cultures.

Wikipedia lists some 2000 controversial issues.

8.1 Task for Controversy Detection

In its simplest form, for the Information Nutrition Label, we can calculate the number of controversial subjects in the text. A more evolved form would to calculate the density of controversial subjects in the text.

8.2 Methods for Controversy Detection

One method we can suggest for calculating the controversy of a text would be to look at those papers that implement Wikipedia featured article detection: they have to address the controversy flaw (the developed technology has parts that apply to non-Wikipedia articles as well). For topics that are covered by Wikipedia, determine the portion of reverts (after article editing), the so-called edit wars in Wikipedia. See the coverage measure (essay articles) below. Compute a number of features that hint controversy: topicality, retweet number and probability, query logs.

8.3 Data sets for Controversy Detection

Data Sources: Aharoni *at al.* (see further reading) describes a novel and unique argumentative structure dataset. This corpus consists of data extracted from hundreds of Wikipedia articles using a meticulously monitored manual annotation process. The result is 2,683 argument elements, collected in the context of 33 controversial topics, organized under a simple claim-evidence structure. The obtained data are publicly available for academic research.

The paper by Dori-Hacohen and Allan below also has a data set.

Test cases

Balance the number of pro and con arguments, using an argument search engine[43]) For queries/documents, which contain one of the controversial topics listed on the Wikipedia page, search/find documents that discuss (essay-like style) a topic. Choose documents appropriate for a specific reading level/background. Extract keywords/concepts and measure the overlap with controversial topics list (Wikipedia), debate portals, and the like.

8.4 Further reading for Controversy Detection

1. Wikipedia "Controversy" article [44]
2. Wikipedia list of controversial issues [45]

[43] ttp://141.54.172.105:8081/
[44] https://en.wikipedia.org/wiki/Controversy
[45] https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues

3. Examples of discussions of controversial topics can be found in the Scientific American[46] and on Plato[47].

4. Aharoni, Ehud, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. "A Benchmark Dataset for Automatic Detection of Claims and Evidence in the Context of Controversial Topics." In ArgMining@ACL, pp. 64-68. 2014.

5. Dori-Hacohen, Shiri, and James Allan. "Detecting controversy on the web." In Proceedings of the 22nd ACM international conference on Conference on information knowledge management, pp. 1845-1848. ACM, 2013. "... Our approach maps a webpage to a set of Wikipedia articles, and uses the controversiality of those ... used two stop sets, the 418 INQUERY stop set [4] or a short, 35 term set (Full vs. ... 3. Handling non labeled data: We use two alterna- tives to fill in the blanks when labeled data ..."

9 Authority / Credibility / Trust

For the Information Nutrition Label, we consider trust and authority as synonyms that refer to a property of the source of a message, while credibility is an attribute of the message itself. On the Web, trust is assigned to a web site, while the different pages of the web site may be different in terms of credibility. When looking at a single document, users are most interested in its credibility; on the other hand, even experienced users judge credibility mainly based on their trust of the source. In the same way, for a system, however, it is easier to estimate the authority of a source (based on the information available), while there might be little document-specific evidence concerning its credibility.

9.1 Task for Authority

The task is to determine the authority or trust of the source of a document. Here we focus on Web sites and social media as sources.

9.2 Methods for Authority

For Web sites, a large number of methods for estimating authority have been proposed, of which we mention just a few:

- PageRank (Further Reading 1) is the most popular method for computing the importance of a Web site.
- Kleinberg's HITS algorithm (Further Reading 2) distinguishes between hub and authority scores.
- BrowseRank (Further Reading 3) computes the importance of a Web site by analysing user behavior data.

[46] https://www.scientificamerican.com/article/fact-or-fiction-vaccines-are-dangerous/
[47] https://plato.stanford.edu/entries/creationism/

- Alexa Rank[48] measures Web site's popularity based solely on traffic to that site, in the form of a combined measure of unique visitors and page views of a website.

Recently, there also have been some approaches addressing the credibility of social media messages:

- Tweetcreed [4] is a Chrome browser extension computing a credibility score for a tweet using six types of features: meta-data, content-based simple lexical features, content-based linguistic features, author, external link URL's reputation, and author network.
- Sharriff et al. (Further Reading 5) aimed at estimating credibility perception of Twitter news considering features such as reader demographics, news attributes and tweet features.
- Popat et al. (Further Reading 6) presents a method for automatically assessing the credibility of claims in a message, which retrieves corresponding articles and models their properties such as the stance language style, their reliability, time information as well as their interrelationships.

9.3 Data sets for Authority and Trust

- Kakol et al. (Further Reading 7) provides a manually annotated dataset that can be used for credibility prediction [49].
- Popat et al. (Further Reading 6) collected data from Wikipedia and snopes.com[50].

9.4 Further reading for Authority and Trust

1. Page, L., Brin, S., Motwani, R., Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Stanford InfoLab.
2. J. Kleinberg. Authoritative sources in a hyperlinked environment. J. ACM, 46:604632, 1999.
3. Yuting Liu , Bin Gao , Tie-Yan Liu , Ying Zhang , Zhiming Ma , Shuyuan He , Hang Li, BrowseRank: letting web users vote for page importance, Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, July 20-24, 2008, Singapore, Singapore [doi¿10.1145/1390334.1390412]
4. Gupta, Aditi, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. ”Tweetcred: A real-time Web-based system for assessing credibility of content on Twitter.” In Proc. 6th International Conference on Social Informatics (SocInfo). Barcelona, Spain. 2014.
5. Shafiza Mohd Shariff, Xiuzhen Zhang, Mark Sanderson. “On the credibility perception of news on Twitter: Readers, topics and features.” Computers in Human Behavior 75 (2017) 785-794.

[48] https://www.alexa.com
[49] https://github.com/s8811/reconcile-tags
[50] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/impact/web-credibility-analysis/

6. Kashyap Popat, Subhabrata Mukherjee, Jannik Strtgen, and Gerhard Weikum. ”Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media.” In Proceedings of the 26th International Conference on World Wide Web Companion, pp. 1003-1012. International World Wide Web Conferences Steering Committee, 2017.

7. Kakol, Michal, Radoslaw Nielek, and Adam Wierzbicki. “Understanding and predicting Web content credibility using the Content Credibility Corpus.” Information Processing Management 53, no. 5 (2017): 1043-1061.

10 Technicality

An article may be well written and grammatically understandable, but its content may cover concepts only understandable to people learned in a certain domain. These documents may deal with a technical issue or use a large proportion of technical terms.

10.1 Task for Technicality Measurement

For our Information Nutrition Label, we want to calculate a technicalness score, or technicality, for a document that indicates how hard it would be to understand for someone outside the field.

10.2 Methods for Technicality Measurement

Similar to Readability, but more related to content than form, Technicality is a property of a document capturing the proportion of the domain-specific vocabulary used by the document. Style-based features are already captured by the readability score.

10.3 Data sets for Technicality Measurement

Data Sources:

- Terminology extraction software [51][52]
- Further tools are available from [53]
- In Wikipedia, external links provide a set of freely available tools under ”Terminology Extraction”[54]
- Word frequency information[55] (English), in German[56][57][58], in other languages[59]

[51] https://github.com/termsuite/termsuite-core
[52] https://github.com/texta-tk/texta
[53] https://github.com/search?utf8=%E2%9C%93&q=terminology+extraction
[54] https://en.wikipedia.org/wiki/Terminology_extraction
[55] http://www.wordfrequency.info
[56] http://corpus.leeds.ac.uk/frqc/internet-de-forms.num
[57] http://www1.ids-mannheim.de/kl/projekte/methoden/derewo.html
[58] http://wortschatz.uni-leipzig.de/de/download
[59] https://web.archive.org/web/*/http://wortschatz.uni-leipzig.de/html/wliste.html

Test cases and benchmarks:

- ACL RD-TEC[60]. QasemiZadeh, Behrang, and Anne-Kathrin Schumann. "The ACL RD-TEC 2.0: A Language Resource for Evaluating Term Extraction and Entity Recognition Methods." In LREC. 2016.
- GENIA Corpus[61] is a popular corpus that has been used to evaluate various ATE algorithm for the last decade. In JATE2, instead of using the annotation file "GENIAcorpus302.xml", the 'concept.txt' containing a breakdown list of GENIA concepts and relations (more like ontology) are used as the "Gold Standard" (GS) list.

10.4 Further reading for Technicality Measurement

1. Justeson, John S., and Slava M. Katz. "Technical terminology: some linguistic properties and an algorithm for identification in text." Natural language engineering 1, no. 1 (1995): 9-27.
2. Dagan, Ido, and Ken Church. "Termight: Identifying and translating technical terminology." In Proceedings of the fourth conference on Applied natural language processing, pp. 34-40. Association for Computational Linguistics, 1994.
3. Pazienza, Maria, Marco Pennacchiotti, and Fabio Zanzotto. "Terminology extraction: an analysis of linguistic and statistical approaches." Knowledge mining (2005): 255-279.

11 Topicality

Topical documents are documents which cover topics that are in the current zeitgeist. This measure, on the information label, will be a time-dependent value. That is, an information label with this value one month, might have a different value the next.

11.1 Task for Topicality Detection

Topicality detection here means to decide whether the document is of current interest or not. One of the salient points of the negative effect of fake news was to falsely influence thinking about things in the current news cycle.

11.2 Methods for Topicality Detection

Extract the salient terms (keyterms) and entities of the document. Compare those terms to the terms found in recent news or publications, or search engine queries.

Tools:

- Text Mining Online[62]

[60] https://github.com/languagerecipes/acl-rd-tec-2.0

[61] https://github.com/ziqizhang/jate/wiki/Evaluation-and-Dataset

[62] http://textminingonline.com/how-to-use-stanford-named-entity-recognizer-ner-in-python-nltk-and-other-programming-languages Keyphrase extraction

- KeyPhrase Extraction[63][64]

11.3 Data sets for Topicality Detection

Current topics can be found on these sites, for example, ABC News[65], or lists of current events[66]. Current news and compiled multilingual lists of entities can be found at the UE-funded EMM NewsExplorer[67]

11.4 Further reading for Topicality Detection

1. Zafar, Muhammad Bilal, et al. ”Zafar, Muhammad Bilal, Parantapa Bhattacharya, Niloy Ganguly, Saptarshi Ghosh, and Krishna P. Gummadi. ”On the Wisdom of Experts vs. Crowds: Discovering Trustworthy Topical News in Microblogs.” In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work Social Computing, pp. 438-451. ACM, 2016
2. Wu, Baoning, Vinay Goel, and Brian D. Davison. ”TWu, Baoning, Vinay Goel, and Brian D. Davison. ”Topical trustrank: Using topicality to combat web spam.” In Proceedings of the 15th international conference on World Wide Web, pp. 63-72. ACM, 2006.
3. Diakopoulos, Nicholas, and Arkaitz Zubiaga. ”Newsworthiness and Network Gatekeeping on Twitter: The Role of Social Deviance.” In ICWSM. 2014.

Acknowledgements

This proposition is the result of a workshop held during the Dagstuhl Seminar number 17301 on User-Generated Content in Social Media, July 23 28, 2017.

References

[1] A. K. Al Tamimi, M. Jaradat, N. Al-Jarrah, and S. Ghanem. Aari: automatic arabic readability index. *Int. Arab J. Inf. Technol.*, 11(4):370–378, 2014.

[2] M. Anderka. *Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia*. Dissertation, Bauhaus-Universität Weimar, June 2013.

[3] M. Anderka, B. Stein, and N. Lipka. Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. In B. Hersh, J. Callan, Y. Maarek, and M. Sanderson, editors, *35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12)*, pages 981–990. ACM, Aug. 2012.

[63] https://github.com/luffycodes/KeyphraseExtraction , https://github.com/Gelembjuk/keyphrases
[64] https://github.com/snkim/AutomaticKeyphraseExtraction
[65] http://abcnews.go.com/topics/
[66] http://libguides.umflint.edu/topics/current or http://www.libraryspot.com/features/currentevents.htm
[67] http://emm.newsexplorer.eu/NewsExplorer/home/en/latest.html

[4] A. Balahur, J. M. Hermida, and A. Montoyo. Detecting implicit expressions of emotion in text: A comparative analysis. *Decision Support Systems*, 53(4):742–753, 2012.

[5] M. L. Bernard, B. S. Chaparro, M. M. Mills, and C. G. Halcomb. Comparing the effects of text size and format on the readibility of computer-displayed times new roman and arial text. *International Journal of Human-Computer Studies*, 59(6):823–835, 2003.

[6] B. Broda, B. Niton, W. Gruszczynski, and M. Ogrodniczuk. Measuring readability of polish texts: Baseline experiments. In *LREC*, pages 573–580, 2014.

[7] K. Collins-Thompson. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135, 2014.

[8] S. A. Crossley, J. Greenfield, and D. S. McNamara. Assessing text readability using cognitively based indices. *Tesol Quarterly*, 42(3):475–493, 2008.

[9] S. A. Crossley, J. Greenfield, and D. S. McNamara. Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42(3):475–493, 2008.

[10] O. De Clercq, V. Hoste, B. Desmet, P. Van Oosten, M. De Cock, and L. Macken. Using the crowd for readability prediction. *Natural Language Engineering*, 20(3):293–325, 2014.

[11] L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284. Association for Computational Linguistics, 2010.

[12] T. François. An analysis of a french as a foreign language corpus for readability assessment. In *Proceedings of the third workshop on NLP for computer-assisted language learning at SLTC 2014, Uppsala University*, number 107. Linköping University Electronic Press, 2014.

[13] R. H. Hall and P. Hanna. The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. *Behaviour & information technology*, 23(3):183–195, 2004.

[14] K. Lee, Y. Artzi, Y. Choi, and L. Zettlemoyer. Event detection and factuality assessment with non-expert supervision. 2015.

[15] M. Levis, M. Helfert, and M. Brady. Information quality management: Review of an evolving research area. 01 2007.

[16] W.-H. Lin, T. Wilson, J. Wiebe, and A. Hauptmann. Which side are you on?: identifying perspectives at the document and sentence levels. In *Proceedings of the tenth conference on computational natural language learning*, pages 109–116. Association for Computational Linguistics, 2006.

[17] A. Lotan, A. Stern, and I. Dagan. Truthteller: Annotating predicate truth. 2013.

[18] A.-L. Minard, M. Speranza, R. Urizar, B. Altuna, M. van Erp, A. Schoen, C. van Son, et al. Meantime, the newsreader multilingual event and time corpus. 2016.

[19] E. Pitler and A. Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the conference on empirical methods in natural language processing*, pages 186–195. Association for Computational Linguistics, 2008.

[20] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein. A stylometric inquiry into hyperpartisan and fake news. *CoRR*, abs/1702.05638, 2017.

[21] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky. Linguistic models for analyzing and detecting biased language. In *ACL (1)*, pages 1650–1659, 2013.

[22] R. Saurí and J. Pustejovsky. Factbank: a corpus annotated with event factuality. *Language resources and evaluation*, 43(3):227, 2009.

[23] R. Saurí and J. Pustejovsky. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299, 2012.

[24] S. Stajner and H. Saggion. Readability indices for automatic evaluation of text simplification systems: A feasibility study for spanish. In *IJCNLP*, pages 374–382, 2013.

[25] G. Stanovsky, J. Eckle-Kohler, Y. Puzikov, I. Dagan, and I. Gurevych. Integrating deep linguistic features in factuality prediction over unified datasets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 352–357, 2017.

[26] S. Stieglitz and L. Dang-Xuan. Emotions and information diffusion in social mediasentiment of microblogs and sharing behavior. *Journal of Management Information Systems*, 29(4):217–248, 2013.

[27] R. Wang and D. Strong. Beyond accuracy: what data quality means to data consumers. *Journal of management information systems*, 12(4):5–33, 1996.

[28] B. L. Zakaluk and S. J. Samuels. *Readability: Its Past, Present, and Future.* ERIC, 1988.