A Study of Factuality, Objectivity and Relevance: Three Desiderata in Large-Scale Information Retrieval?

Christina Lioma
Department of Computer Science
University of Copenhagen
Universitetsparken 5, 2100 Copenhagen,
Denmark
c.lioma@di.ku.dk

Wei Lu
School of Information Management
Wuhan University
No. 299 Bayi Road, Wuhan, China
weilu@whu.edu.cn

Birger Larsen
Department of Communication
Aalborg University in Copenhagen
A.C. Meyers Vaenge 15, 2450 Copenhagen,
Denmark
birger@hum.aau.dk

Yong Huang
School of Information Management
Wuhan University
No. 299 Bayi Road, Wuhan, China
huangyng@gmail.com

ABSTRACT

Much of the information processed by Information Retrieval (IR) systems is unreliable, biased, and generally untrustworthy [15, 45, 48]. Yet, factuality & objectivity detection is not a standard component of IR systems, even though it has been possible in Natural Language Processing (NLP) in the last decade. Motivated by this, we ask if and how factuality & objectivity detection may benefit IR. We answer this in two parts. First, we use state-of-the-art NLP to compute the probability of document factuality & objectivity in two TREC collections, and analyse its relation to document relevance. We find that factuality is strongly and positively correlated to document relevance, but objectivity is not. Second, we study the impact of factuality & objectivity to retrieval effectiveness by treating them as query independent features that we combine with a competitive language modelling baseline. Experiments with 450 TREC queries show that factuality improves precision by more than 10% over strong baselines, especially for the type of uncurated data typically used in web search; objectivity gives mixed results. An overall clear trend is that document factuality & objectivity is much more beneficial to IR when searching uncurated (e.g. web) documents vs. curated (e.g. state documentation and newswire articles).

To our knowledge, this is the first study of factuality & objectivity for back-end IR, contributing novel findings about the relation between relevance and factuality/objectivity, and statistically significant gains to retrieval effectiveness in the competitive web search task.

CCS Concepts

•**Information systems** → **Content analysis and feature selection;** *Retrieval models and ranking;*

Keywords

information retrieval; natural language processing; large-scale content analysis

1. INTRODUCTION

Information *factuality* is typically understood as the extent to which accounts, opinions, evaluations or emotions that vary across multiple reporters are factual [17]. Information is factual, hence reliable, if one can accept it as the truth without needing to check its validity elsewhere [56]. Information *objectivity* on the other hand, is almost always defined as the opposite of subjectivity, which is generally understood as the semantic orientation or valence of text [21, 50], or more simply the extent to which text meaning depends on the author's perspective [23]. Objectivity is the degree to which language is interpretable independently of the speaker's perspective [32]. Factuality and objectivity are well established cognitive and linguistic notions [23], both often manifested when opinions or evaluations are conveyed.

Intuitively, retrieving unreliable and subjective information seems undesirable, as it can impact negatively decision making [17] and may require time-consuming fact cross-validation from other sources [56]. However, whereas assessing factuality & objectivity seems a standard cognitive task for humans [8], search engine users often mistake the output of web search engines as a confirmation of factuality or objectivity. For instance, many interpret the ranking of search results as a key indicator of credibility [20], whereas social network users often determine the truthfulness of content based on misleading heuristics, such as if an item is retweeted [48]. In fact, in 2005, a majority of U.S. web users considered search engines fair and unbiased as an information source [15]. This is not always the case; much of the information on the web is actually unreliable, biased, and generally untrustworthy [15, 45, 48].

As a result of this, IR research has started addressing

aspects of factuality & objectivity (reviewed in Section 2). NLP research has been addressing the detection and measurement of factuality & objectivity for a longer time, with significant recent advances and resulting gains for tasks like question answering [49] or information extraction [60].

Motivated by the need to distinguish between retrieving credible over non-credible information, and by the NLP advances that allow its detection, we ask if and how document factuality & objectivity can benefit IR. We choose factuality & objectivity because they are two core dimensions of credibility [17, 50]. We use two state-of-the-art, scalable NLP methods to estimate document factuality & objectivity in two TREC collections (Section 3). Statistical analysis of the collections (Section 4) and subsequent experimental evaluation in a reranking scenario (Section 5) show that (i) considering the factuality of the retrieved documents improves retrieval effectiveness significantly for uncurated (e.g. web) collections but not for curated collections (e.g. older and 'cleaner' TREC disks); (ii) document objectivity does not seem to improve retrieval effectiveness. This work contributes novel findings about the relation between relevance and factuality/objectivity, and statistically significant gains to retrieval effectiveness in the competitive web search task.

2. RELATED WORK

Previous research has investigated aspects of information factuality & objectivity in search scenarios. One example is Dispute Finder [12], a system that identifies contentious topics by looking for disputed text: a factual claim is considered disputed if a webpage suggests both that the claim is false and also that other people say it is true. Dispute Finder extracts disputed claims by searching the web for patterns such as `falsely claimed that X` and using them to train a classifier to select text that makes a disputed claim. Similar to this is the *factual density* measure [35], which estimates factuality as the ratio of facts in a document over document length. *Factual density* has been used to identify highly factual (hence reliable) articles in Wikipedia [35] and other web sources [24]. Visualisations on the user interface have also been shown to help users assess information as factual, biased or opinionated [18, 48, 51]. E.g., visualising the Wikipedia edit history can change users' perceptions of trustworthiness [30]; similarly, visualizing popularity among experts next to the search results can help users make more accurate credibility ratings [56].

Several text quality metrics exist for capturing aspects of factuality & objectivity, typically using as features stylometric indicators of text quality [34], for instance character trigram distributions [42], word counts [5], or text structure (for blogs) [26]. Some approaches are so far applied solely to curated data, like Wikipedia [34, 42], which is relatively homogeneous (format- and structure-wise). Processing uncurated web data requires more complex features, such as semantic relations [13], extracted or inferred about entities named in text [35], using as clues individual words and/or lexico-syntactic patterns [60]. The above approaches are developed stand-alone, not as an integral part of IR systems. However, methods for potentially integrating such text quality metrics into IR systems abound, for instance when text quality is interpreted as ratios of (combinations of) stopwords over content words per document [4, 64, 65]; term length [27] or part-of-speech (for ranking [38, 41], but also for index pruning [38]); technical [33] or ambiguous scientific

terminology [36]; non-compositional multiword expressions [39, 47]; document readability [9, 19, 29, 43, 44]; document discourse [37] or coherence [52, 40].

Factuality and objectivity analysis has been studied in NLP for over a decade, producing annotated corpora [61], schemes for manual annotation [22], tools for automatic analysis [62], and generally a good understanding of their computational treatment [58]. This has led to successful applications of factuality & objectivity analysis to sentiment analysis [63], information extraction [60], and question answering [49], in English but also other languages [2]. In these tasks, the analysis is typically done at a phrase, sentence, or document level; however, more recent analysis is done at a lexical level, e.g. to disambiguate words according to their usage or sense [2], or for lexical semantics research *per se*, e.g. showing that objective adjectives are most likely to modify concrete nouns [23].

In this work, we measure factuality & objectivity on a document level, so that we can study its relation and potential usefulness to retrieval. None of the above methods have been applied to back-end retrieval to our knowledge.

3. FACTUALITY AND OBJECTIVITY ESTIMATION

The first step in studying factuality & objectivity for IR is detecting them in documents. We do this using two state-of-the-art, scalable approaches. We present these next and their application to IR documents in Section 4.

3.1 Detecting Factuality

We estimate document factuality using *factual density* [35]. This consists of (i) identifying factual and non-factual documents in some small corpus, (ii) using features of these documents to train a classifier, and (iii) applying the classifier to our data to estimate the factuality of each document. *Factual density* uses open information extraction (Open IE) methods to extract relational tuples of facts from text. A relational tuple is typically a triplet of two noun phrase arguments and a relational verb phrase, e.g. f = (Poe, was born in, Boston), where f denotes a fact. Even though Open IE is scalable to large sets of relations and corpora [13], it is prone to extracting noise [14], risking to overestimate *factual density*. To avoid this, we extract facts using ReVerb, a competitive IE system that uses part-of-speech (POS) and lexical constraints on verb relations, with a lower noise extraction rate in web documents [14]. For implementation details see [14].

Following [35], we collect randomly 2000 *featured*[1] or *good*[2] Wikipedia articles (as positive examples of factuality), and 2000 articles from the remaining Wikipedia (as negative examples of factuality). *Featured* and *good* articles are the best articles in Wikipedia, as decided by editorial review according to accuracy, neutrality, completeness, and style, so they are suitable examples of factuality. Even though there is no guarantee that there will not be factual documents among our 2000 negative examples, it is likely that negative examples will be *less* factual than *featured/good* articles. This premise worked well for [35] and we therefore also adopt it. We extract facts from the 4000 positive and

_____

[1] http://en.wikipedia.org/wiki/Wikipedia:Featured_articles
[2] http://en.wikipedia.org/wiki/Wikipedia:Good_articles

Table 1: Syntactic templates for extraction patterns used in objectivity detection (from [60]): *subj* (subject), *dobj* (direct object), *np* (noun phrase), *aux* (auxiliary verb), *prep* (preposition).

[subj] passive-verb	active-verb [dobj]	noun prep [np]
[subj] active-verb	infinitive [dobj]	active-verb prep [np]
[subj] active-verb dobj	verb infitive [dobj]	passive-verb prep [np]
[subj] verb infinitive	noun aux [dobj]	infitive prep [np]
[subj] aux noun		

Table 2: Datasets. # means 'number of'.

	CURATED	UNCURATED
	Disks4-5	ClueWeb09B
# Documents	556,077	50,220,423
# Queries	301-450	1-200
# TREC-assessed Documents	144,144 (25.9%)	83,163 (0.17%)
TREC track	AdHoc	Web AdHoc

negative examples using ReVerb. In 60 of the negative examples, we find no facts; this supports the assumption that *non-featured/non-good* articles are likely to be less factual. We use the word count, fact count, fact density (ratio of fact count over word count), and top 10% information gain relational tuples (600 relational tuples) to train a classifier (SVM, RBF kernel) using LIBSVM and 5-fold cross validation. This yields 90.3% accuracy. Using this trained model on our data outputs a probability of factuality for each document. We use this probability as is, without calibrating it. For potentially more refined results, we could have used any among several probability calibration methods specifically designed for SVM classifiers, based e.g. on sigmoid [54], or Bayesian transformations [31]. We do not do so, given the high classification accuracy obtained, and also to have more control over the effect of factuality on retrieval; extra smoothing might fetch even better results, but with the risk of washing away other effects that might make our interpretation of the impact of factuality on relevance less transparent.

3.2 Detecting Objectivity

We estimate document objectivity with the subjectivity detection approach of Wiebe & Riloff [60]. The required inputs are a subjectivity lexicon, a small set of seed nouns, and some human review, which we explain below. For implementation details see [60].

We split our data into sentences with openNLP, and parse sentences to get typed dependencies (e.g. what is the subject or object of the sentence) with the Stanford parser[3]. We filter the sentences using the extraction pattern templates of [60] shown in Table 1, and keep only sentences containing these patterns. We use the top 20 most subjective nouns in our data (manually compiled by examining the words with the highest collection term frequency that are nouns and choosing the 20 most subjective) as seeds and run 400 iterations of Meta-Bootstrapping [55] and Basilisk [59] on the extracted sentences. We select five new subjective nouns after each iteration and manually annotate the resulting 4000 nouns as strongly/weakly subjective or objective. We add these annotated nouns to an existing subjectivity lexicon of >8000 entries made available in [60], and use them as features to classify sentences as subjective or objective in our dataset. A sentence is subjective if it contains ≥ 2 strongly subjective terms, and objective if it contains no strongly subjective terms and maximum 2 weakly subjective terms (as per [60]). We thus build 10000 training examples (5000 subjective and 5000 objective), which we feed into an extraction pattern learning algorithm that extracts subjective and objective patterns, such as [*subj*] complained, or to condemn [*dobj*] (see Table 1 for abbreviations). We rank the extracted

---

[3]http://nlp.stanford.edu/software/lex-parser.shtml

patterns according to the probability that a sentence is subjective or objective given that a specific pattern appears in it, as described in [60], using the exact same thresholds.

We use the above subjective/objective nouns, lexicon, patterns, plus POS features, to train an objectivity classifier (SVM, RBF kernel) using LIBSVM and 5-fold cross validation, on our 10000 training sentences. This yields 87.83% accuracy. We use this trained model to estimate the probability of objectivity of each document in our data as a fraction of the number of objective sentences in it over its total number of sentences.

4. STATISTICAL ANALYSIS OF FACTUAL-ITY & OBJECTIVITY VS. RELEVANCE

To find out if factuality & objectivity are useful to IR we study their relation to relevance. We start with a small-scale analysis of 227,307 documents, for which we have scores of relevance, factuality & objectivity (explained below).

We look at the association between relevance and factuality/objectivity on two datasets of different domains: (a) a curated, 'clean' newswire/official documentation corpus (TREC Disks4-5), and (b) a bigger, uncurated, relatively noisy web crawl, TREC ClueWeb09B (see Table 2). Disks4-5 contain documents from the 103rd Congressional Record, the 1994 Federal Register, the 1992-1994 Financial Times, the 1996 Foreign Broadcast Information Service, and the 1989-1990 Los Angeles Times. We consider these curated, because they have been very likely professionally edited. ClueWeb09B contains webpages crawled in 2009 from heterogeneous web sources that are not necessarily professionally edited and include spam; hence, we consider this data uncurated. We select these datasets intentionally to study the effect of curated vs. uncurated data on factuality & objectivity detection for IR. Both datasets come with TREC queries and binary relevance assessments. We study the relation between relevance and factuality/objectivity on a subset of these datasets which includes *all and only* those documents for which we have TREC relevance assessments: these are ∼144K documents or 25.9% of Disks4-5, and ∼83K documents or 0.17% of ClueWeb09B (see Table 2). We estimate the factuality & objectivity of each document in these subsets as described in Section 3, and analyse their distribution (Sec 4.1) and their covariation with document relevance (Section 4.2).

4.1 Factuality & Objectivity Distribution

We start our analysis by looking at how factuality & objectivity are distributed. Knowing how their scores are spread out is important for using them in IR, e.g. to avoid bias or saturation effects that may result from heavily skewed distributions [53]. To get a clearer visualisation of trends, we use binning. We sort all documents by their probability of factuality & objectivity, and we divide them into bins. We
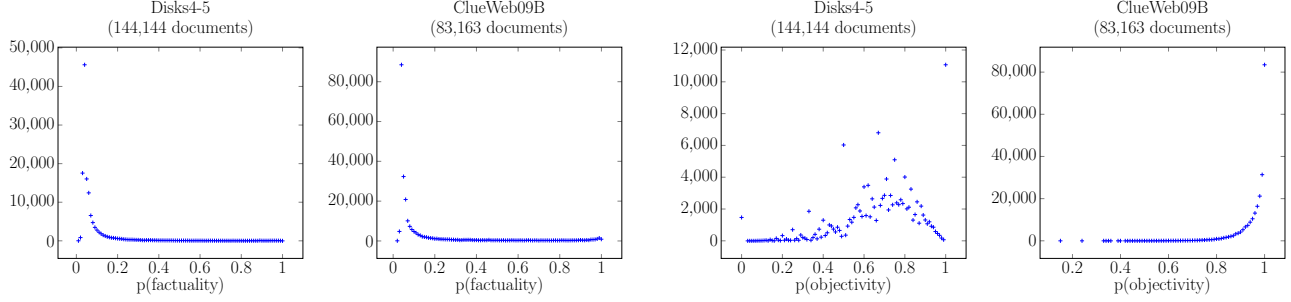
Figure 1: Distribution of probability of factuality/objectivity (binned). The y-axis shows the number of documents.
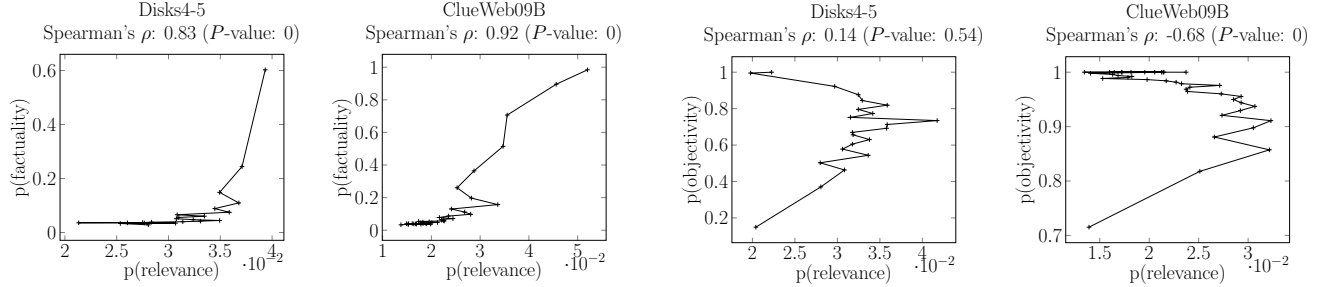

Figure 2: Probability of factuality & objectivity (y-axis) vs. probability of relevance (x-axis) (binned).

estimate the number of bins using Scott's formula [57]:

$$M = \frac{R}{3.49s}N^{1/3} \tag{1}$$

where M is the number of bins, R is the range, N is the number of data points, and s is the sample variance. This results in: (a) 23 equal-sized bins of 6000 documents and 1 bin of the remaining 6144 documents for Disks4-5; (b) 40 equal-sized bins of 2000 documents and 1 bin of the remaining 3163 documents for ClueWeb09B. The probability of factuality & objectivity in each bin is the mean of the factuality & objectivity probabilities in that bin. The frequency of a probability in a bin is the sum of the frequencies of the probabilities in that bin.

Figure 1 shows the distribution of the probabilities of factuality & objectivity in our data. We see a heavy-tailed trend for factuality: a very large number of documents have a very low (but nonzero, namely 0.04) probability of factuality, and only very few have high factuality. This trend is the same in both datasets, but the peak of the distribution is at a much higher value in ClueWeb09B (~80K) than in Disks4-5 (~ 45K). This means that the number of low factuality documents is much higher in ClueWeb09B (~80K/83K ≈ 94% of all documents) than in Disks4-5 (~45K/144K ≈ 32% of all documents). This may be due to the curated vs. uncurated domain difference between the collections: official or journalistic documents are overall more likely to be reliable than documents crawled from the web.

The probability of objectivity (Figure 1) is distributed almost inversely to factuality: there seem to be more documents with higher than lower objectivity, implying that most documents are not highly biased. Note again the difference in scale of the peak of the distribution for the two datasets: the number of highly objective documents is ~11K/114K

≈ 8% of Disks4-5, but ~80K/83K ≈ 94% of ClueWeb09B. For Disks4-5 it is not surprising that only 8% of all official/journalistic documentation is very objective, while the remaining documents have varying degrees of objectivity, without a heavy tail. Official state documentation tends to be void of personal opinions and fairly impartial, but journalistic publications, which make up most of Disks4-5, often include subjectivity [1]; assuming that $p(objectivity) < 0.5$ implies subjectivity, this corresponds to 0.0-0.4 on the x-axis. For ClueWeb09B however, it is surprising to see such a high rate (94%) of very objective documents. This does not seem very intuitive and could be due to misestimations of the objectivity detection. To understand the causes of this, we look at the presence of spam in the ClueWeb09B subset.

The effect of spam in ClueWeb09B.

We use the spam scores of [10] for ClueWeb09B, where scores >70 are interpreted as spam. Under this interpretation, spam makes up ~70% of the whole ClueWeb09 and ~49% of our TREC-assessed subset of ClueWeb09B. Figure 3 plots the probability of factuality & objectivity in the ClueWeb09B subset vs. the spam score per bin. The same binning described earlier is used. The spam score of each bin is the mean of all spam scores in that bin. In Figure 3, left plot, we see an almost linear relation between factuality and spam, where higher factuality corresponds to lower spam, and vice versa. This makes sense intuitively; spam is notoriously unreliable. Even though the correlation between factuality and spam is not extremely strong (Spearman's ρ: 0.6), it is statistically significant, and the trend is visually fairly clear. The dense concentration of spam (~50-60 on the y-axis) corresponds to a low probability of factuality (<0.2). This indicates that most of the documents are spam
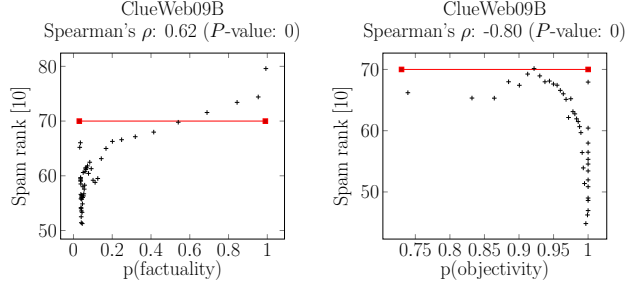
Figure 3: Probability of factuality & objectivity (x-axis) vs. spam rank [10] (y-axis), (binned). Points below the horizontal line are considered spam.

and of low factuality, which agrees with our earlier finding that ~49% of this data is spam.

For objectivity (Figure 3, right plot) we see a different trend: there is a strong negative correlation between objectivity and spam (Spearman's ρ: –0.8), and the relation is not diagonally linear but rather mostly 'vertical', i.e. documents of high objectivity have a wide range of spam scores. Furthermore, there are very few non-spam documents: only one bin is just above the spam threshold, i.e. 2000 documents ≈ 2% of the whole subset. This 2% of non-spam tends to have high objectivity (~0.9), which seems intuitively reasonable. The remaining 98% of the documents are spam and have various degrees of objectivity (all however are fairly highly objective, namely > 0.7 on the x-axis, as pointed out above). Therefore most of the ~80K highly objective documents in ClueWeb09B shown in Figure 1 must be spam, because only ~2000K of them can be non-spam.

The strong presence of spam in the TREC-assessed subset of ClueWeb09B may impact objectivity detection more than factuality detection. The cause could be that spam contains *word stuffing*: gratuitous keywords inserted to improve the retrieved rank of the document [10]. Whereas our detection of factuality relies mainly on relations denoted by verbs (relational tuples), our detection of objectivity, while also using verb patterns, emphasises primarily noun or noun phrase lists associated with objectivity & subjectivity (see Section 3). *Word stuffing* is typically manifested by the insertion of nouns and adjectives rather than verbs. Hence, it is likely that word stuffing may muddle the nominal processing of objectivity detection in ClueWeb09B, and result in artificially inflated objectivity scores. Manual inspection of the objective nouns and lexicon used for our objectivity detection confirms that this is the case.

4.2 Factuality, Objectivity and Relevance

We compare the above probabilities of factuality & objectivity of our data to their probability of relevance: given the document bins of factuality & objectivity described in Section 4.1, we estimate the probability that a randomly selected relevant document belongs to a bin as:

$$p(d \in b_i | relevance) = \frac{|d_{RLV} \in b_i|}{|d_{RLV}|} \quad (2)$$

where d_{RLV} is a relevant document according to the TREC relevance assessments, b_i is the ith bin, and $|\cdot|$ denotes cardinality. We refer to $p(d \in b_i | relevance)$ as the probability of relevance in that bin.

Figure 2 plots the probability of relevance (x-axis) vs. the probability of factuality or objectivity per bin (y-axis). We see that $p(relevance)$ varies non-randomly across bins: factuality is positively, strongly and almost linearly correlated to relevance (Spearman's ρ: 0.83 for Disks4-5; Spearman's ρ: 0.92 for ClueWeb09B). This means that boosting the ranking of higher factuality documents may boost retrieval performance, especially for ClueWeb09B where the correlation is stronger. The range of $p(factuality)$ is smaller for Disks4-5 (up to ~0.6) than for ClueWeb09B (up to ~1.0). On first thought this seems to indicate that there are documents of higher factuality in ClueWeb09B than in Disks4-5. Closer inspection however reveals that this is due to the different size of bins in Disks4-5 (6000 documents) and ClueWeb09B (2000 documents). The bigger the size of the bin, the harder to get a mean close to the maximum value of 1[4]; it is more likely that the mean is 'diluted' by values away from the maximum, hence it drops.

Objectivity (Figure 2) is not correlated to relevance in Disks4-5 (Spearman's ρ: 0.14), and only negatively and non-linearly correlated to relevance in ClueWeb09B (Spearman's ρ: -0.68). The non-linear shape means that the lower probability of relevance corresponds to documents of both minimum and maximum probability of objectivity. Hence, using these objectivity scores for retrieval may produce mixed results. We also note that there seem to be more high objectivity documents in ClueWeb09B (most points sit above 0.9 on the y-axis) than in Disks4-5. This was already indicated in Figure 1, where we discussed that 98% of these documents are considered spam, meaning that measuring their objectivity is likely biased by spam effects such as *word stuffing*. Hence, using our ClueWeb09B objectivity scores for retrieval is likely to produce inconclusive results.

Overall, the findings of this section are: (I) There is a strong positive correlation between relevance and factuality, especially in ClueWeb09B. It is likely that factuality may be useful for retrieval, and especially for web documents. (II) Objectivity gives a mixed picture: for ClueWeb09B, 98% of documents, when binned by their probability of objectivity, appear to be spam, which renders their processing and subsequent analysis inconclusive; for Disks4-5, no correlation is found at all. Hence, it is likely that objectivity may not be useful for retrieval, at least for uncurated data. (III) There is noise in the above estimations, particularly when data (i) includes spam, and (ii) becomes sparse (relevant documents << factual or objective documents). This affects the accuracy of the above estimations, which we should treat as indications only. It remains to be seen experimentally how indicative these correlations are. We do this next.

5. EXPERIMENTAL EVALUATION OF FACTUALITY & OBJECTIVITY FOR IR

We experimentally study the potential usefulness of factuality & objectivity to retrieval by treating the probability of factuality & objectivity in a document as a type of query independent feature, such as PageRank, that we combine with a query dependent baseline. The main idea is: (i) attach a static weight to each document based on its factuality or objectivity; and (ii) combine this weight with the query dependent baseline score, to give a new score and ranking.

---

[4]Recall that $p(factuality)$ in a bin is the mean of the probabilities of factuality of all documents in that bin.

There are three main approaches to combining such features and a query dependent baseline ranking [11]: (i) rank fusion, (ii) language modelling priors, and (iii) relevance score adjustment. In rank fusion, the baseline and the feature scores are turned into two rankings, which are then fused, e.g. using traditional CombMNZ, voting algorithms, Bayesian inference [3], or more recent rank interleaving methods from online learning [6, 7]. In a language modelling framework, prior probabilities of the feature are calculated and combined with the language modelling probability. The third approach, and the one we use, is to rerank based on a combination of baseline and feature scores. This can be done linearly using raw scores, or even better with a nonlinear transformation of the input feature [11]. We choose to use two types of linear combination because, even though they may not give optimal performance, they make it intuitively easy to interpret the impact of the feature score on the final ranking.

5.1 Baselines and our Methods

Our baseline ranking model is a unigram, query likelihood, Dirichlet-smoothed, language model (DIR). We rerank the top 1000 retrieved documents by the baseline according to their factuality or objectivity scores. Limiting the reranking to the top 1000 is more efficient than reranking all documents with a nonzero baseline score, without making a large difference to system effectiveness [11]. Let S be the baseline ranking score of a document, which in this case is a probability but practically computed as a log for efficiency. Let $p(RLB)$ and $p(OBJ)$ be the probability of factuality & objectivity of a document computed as per Section 3. We use two reranking approaches, (i) simple *linear* combination (Equation 3), and (ii) *satu* [11] (Equation 4), to compute the reranked S of each document, denoted $\widehat{S^F}$ and $\widehat{S^O}$ for factuality- and objectivity-based reranking:

$$\widehat{S^F_{linear}} = \log S \times \alpha + \log p(FCT) \times (1-\alpha) \qquad (3)$$

where $0 \leq \alpha \leq 1$ is a smoothing parameter controlling the effect of S over $p(FCT)$.

$$\widehat{S^F_{satu}} = \log S + w \times \frac{\log p(FCT)}{k + \log p(FCT)} \qquad (4)$$

where w, k are parameters: w is the maximum, approached as $\log p(FCT)$ increases; k is the value of $\log p(FCT)$ where *satu* is $w/2$ [11].

For objectivity-based reranking ($\widehat{S^O}$), we replace $p(FCT)$ by $p(OBJ)$ in Equation 3-4.

5.2 Experimental Setup

We use Indri 5.8 for indexing and retrieval without stemming or stop word removal. We use the two TREC test collections analysed in Section 4 and shown in Table 2: Disks4-5 with queries 301-450 (title only) from the AdHoc track of TREC6-8 (minus the Congressional Record for TREC7-8), and ClueWeb09B with queries 1-200 from the Web AdHoc track of TREC 2009-2012. We evaluate effectiveness with standard early and deep precision measures: Mean Average Precision (MAP); Normalised Discounted Cumulative Gain (NDCG); Binary Preference (BPREF), which, unlike MAP, does not treat non-assessed documents as non-relevant; Precision in top 10 (P@10); and Mean Reciprocal Rank (MRR)

of the first relevant result. We test for statistical significance using the student t-test at 95% confidence levels.

The baselines and our reranking methods include parameters that we tune using 5-fold cross-validation. We report the average of the five test folds. We vary DIR's $\mu \in \{100,500,800,1000,2000,3000,4000,5000,8000,10000\}$ and the reranking parameters $\alpha \in \{0.5..1\}$ in steps of 0.05, and $w \in \{0.5..3.5\}$ in steps of 0.5. We set $k = 1$.

5.3 Results

For **factuality reranking** (upper part of Table 3), baseline performance is generally higher for MAP and NDCG in the curated than uncurated datasets (more mixed picture for BPREF, P@10, MRR). This trend agrees with the performances recorded in past TREC proceedings and is likely due to the different characteristics and difficulties of the respective test collections: web queries (for the uncurated data) tend to be shorter and to have fewer relevance assessments in relation to the collection size, hence they are somewhat harder to satisfy than earlier Disks4-5 queries.

For **uncurated data**, factuality reranking always outperforms the baseline, or equals it when the maximum has been reached by the baseline (e.g. for MRR), with the exception of the linear combination for the Web track 2009 where MRR slightly drops from 1.0 to 0.98. All these improvements are statistically significant and agree with the strong correlation between factuality and relevance found in Section 4.2.

For factuality reranking in **curated data**, we see both gains, which are modest, as well as drops in performance. This could be because curated data is of higher general quality, so the distinction between low and high factuality documents is not as sharp as in uncurated data (hence there is less margin for improvement using factuality). This agrees with the weaker correlation between factuality and relevance found in Section 4.2 for the curated dataset.

Overall, *satu* performs better than *linear* most of the times, which implies that even more refined transformations of the factuality probability can potentially lead to higher performance gains.

For **objectivity reranking** (lower part of Table 3), we see very little gain and occasional performance drops in uncurated data. Given the high amount of spam in incurated data (discussed in Section 4.1), we consider these findings inconclusive, i.e. we refrain from drawing conclusions about the potential usefulness or not of objectivity in uncurated data retrieval. For curated data, objectivity gives hardly any gains at all, which agrees with the lack of correlation found between objectivity and relevance in Section 4.2. We conclude that objectivity reranking, as implemented here, does not seem to help curated data retrieval.

We focus the remaining discussion on factuality and discuss some pertinent aspects of our findings.

5.3.1 Early vs. deep precision

Factuality improvements are overall much higher for early precision (up to +47.2% for P@10 and +105.1% for MRR) than for deep precision (up to +5% for NDCG, +16.7% for MAP, and 23.6% for BPREF). BPREF brings higher gains than MAP, which means that many among the retrieved documents are non-assessed and are treated blindly as non-relevant by MAP (whereas BPREF ignores them). NDCG has the lowest overall gains, probably because it punishes relevant documents at lower ranks more harshly than

Table 3: Retrieval effectiveness of the baseline (DIR), shown in grey, vs. two factuality (FCT) and objectivity (OBJ) based reranking approaches (*linear* marked ⊕, and *satu* marked ⊗). ±% is the difference from the baseline. Bold means ≥ baseline. ‡ marks stat. significance; * marks the best score per track & measure.

FACTUALITY-BASED RERANKING

Method	TREC-6 AdHoc		TREC-7 AdHoc		TREC-8 AdHoc		TREC'09 Web AdHoc		TREC'10 Web AdHoc		TREC'11 Web AdHoc		TREC'12 Web AdHoc	
			CURATED DATA (Disks4-5)						UNCURATED DATA (ClueWeb09B)					
MAP														
DIR	.1721		.1981*		.2053*		.1433		.1063		.0977		.1007	
DIR⊕FCT	**.1728**	+0.4%	.1894	−4.6%	.2040	−0.6%	**.1466‡**	+2.3%	**.1214‡**	+14.2%	**.1056‡***	+8.1%	**.1034‡**	+2.7%
DIR⊗FCT	**.1740***	+1.1%	.1863	−6.3%	.2050	−0.1%	**.1476‡***	+3.0%	**.1221‡***	+14.9%	**.1049‡**	+7.4%	**.1175‡***	+16.7%
NDCG														
DIR	.4009		.4426*		.4454		.3279		.3020		.2560		.2795	
DIR⊕FCT	**.4010**	+0.0%	.4349	−1.8%	.4442	−0.3%	**.3299‡**	+0.6%	**.3175‡**	+5.1%	**.2571‡**	+0.4%	**.2932‡**	+4.9%
DIR⊗FCT	**.4018***	+0.2%	.4318	−2.5%	**.4463***	+0.2%	**.3300‡***	+0.6%	**.3260‡***	+7.9%	**.2603‡***	+1.7%	**.2982‡***	+6.7%
BPREF														
DIR	.1966		.2163*		.2191		.2270		.2322		.1642		.2335	
DIR⊕FCT	.1965	+0.0%	.2086	−3.7%	**.2197**	+0.3%	**.2271‡***	+0.0%	**.2542‡***	+9.5%	**.1825‡**	+11.1%	**.2886‡***	+23.6%
DIR⊗FCT	**.2011***	+2.3%	.2105	−2.8%	**.2207***	+0.7%	**.2270‡**	+0.0%	**.2538‡**	+9.3%	**.1861‡***	+13.3%	**.2834‡**	+21.4%
P@10														
DIR	.3180		.4060		.4000		.9405		.2585		.2400		.1980	
DIR⊕FCT	**.3260***	+2.5%	**.4280***	+5.4%	**.4160***	+4.0%	**.9682‡**	+2.9%	**.3445‡**	+33.3%	**.2460‡***	+2.5%	**.2400‡***	+21.2%
DIR⊗FCT	.3140	−1.3%	**.4280***	+5.4%	**.4120**	+3.0%	**.9723‡***	+3.4%	**.3805‡***	+47.2%	**.2640‡***	+10.0%	**.2280‡**	+15.2%
MRR														
DIR	.5800		.7149*		.6932		1.000*		.3589		.4096		.2057	
DIR⊕FCT	**.6176**	+6.5%	.7021	−1.8%	**.6948***	+0.2%	.9817‡	−1.9%	**.4588‡**	+27.8%	**.4939‡***	+20.6%	**.4218‡***	+105.1%
DIR⊗FCT	**.6400***	+10.3%	.6974	−2.5%	**.6941**	+01%	**1.000‡***	+0.0%	**.5048‡***	+40.7%	**.4841‡**	+18.2%	**.4104‡**	+99.5%

OBJECTIVITY-BASED RERANKING

| Method	TREC-6 AdHoc		TREC-7 AdHoc		TREC-8 AdHoc		TREC'09 Web AdHoc		TREC'10 Web AdHoc		TREC'11 Web AdHoc		TREC'12 Web AdHoc	
			CURATED DATA (Disks4-5)						UNCURATED DATA (ClueWeb09B)					
MAP														
DIR	.1721		.1981*		.2053*		.1433		.1063*		.0977		.1007	
DIR⊕OBJ	.1720	−0.1%	.1856	−6.7%	.1970	−4.2%	**.1438‡***	+0.3%	**.1063‡***	+0.0%	**.0980‡***	+0.3%	**.1014‡**	+0.7%
DIR⊗OBJ	**.1732***	+0.6%	.1864	−6.3%	.2039	−0.7%	**.1437‡**	+0.3%	.1059‡	−0.4%	**.0977‡**	+0.0%	**.1023‡***	+1.6%
NDCG														
DIR	.4009*		.4426*		.4454*		.3279		.3020*		.2560		.2795	
DIR⊕OBJ	.4001	−0.2%	.4331	−2.2%	.4185	−6.4%	**.3283***	+0.1%	**.3020***	+0.0%	**.2560**	+0.0%	.2794	+0.0%
DIR⊗OBJ	**.4008**	+0.0%	.4313	−2.6%	.4447	−0.2%	**.3282**	+0.1%	.3017	−0.1%	**.2561***	+0.0%	**.2813***	+0.6%
BPREF														
DIR	.1966		.2163*		.2191*		.2270		.2322		.1642*		.2335*	
DIR⊕OBJ	.1963	−0.2%	.2087	−3.6%	.2132	−2.8%	**.2273‡***	+0.1%	**.2323‡***	+0.0%	**.1642‡***	+0.0%	.2331‡	−0.2%
DIR⊗OBJ	**.2014***	+2.4%	.2114	−2.3%	.2172	−0.9%	**.2273‡***	+0.1%	.2319‡	−0.1%	.1638‡	−0.2%	.2329‡	−0.3%
P@10														
DIR	.3180		.4060		.4000		.9405*		.2585		.2400		.1980*	
DIR⊕OBJ	**.3200***	+0.6%	**.4080**	+0.5%	**.4120***	+3.0%	.9365‡	−0.4%	**.2675‡***	+3.5%	**.2440‡***	+1.7%	.1940‡	−2.1%
DIR⊗OBJ	.3140	−1.3%	**.4240***	+4.4%	**.4100**	+2.5%	.9365‡	−0.4%	**.2655‡**	+2.7%	.2380‡	−6.7%	.1940‡	−2.1%
MRR														
DIR	.5800		.7149*		.6932*		1.000*		.3589*		.4096*		.2057	
DIR⊕OBJ	.5681	−2.1%	.6921	−3.3%	.6913	−0.3%	**1.000‡***	+0.0%	.3498‡	−2.6%	.4095‡	+0.0%	**.2185‡**	+6.2%
DIR⊗OBJ	**.5863***	+0.1%	.6807	−5.0%	.6909	−0.3%	**1.000‡***	+0.0%	.3492‡	−2.8%	.4019‡	−1.9%	**.2474‡***	+20.3%
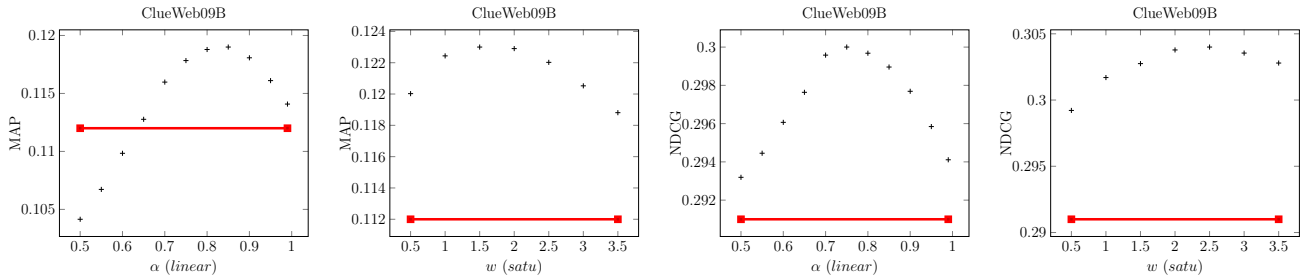
Figure 4: Factuality reranking parameter values for *linear* and *satu* (x-axis) vs. MAP/NDCG (y-axis) in ClueWeb09B. The horizontal line marks baseline performance.
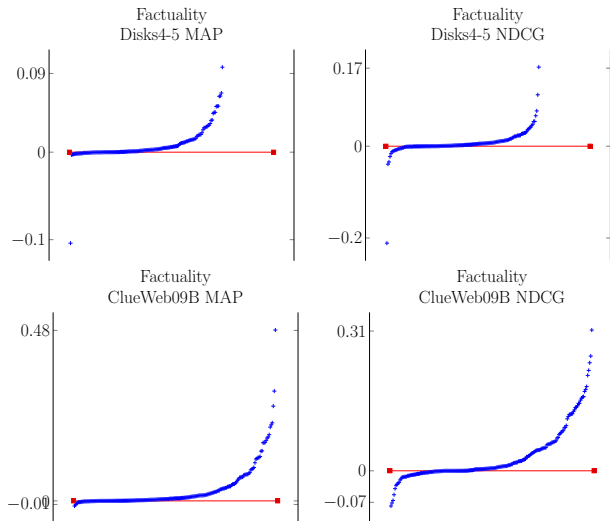
Figure 5: Sorted per-query difference in MAP/NDCG between the baseline (DIR) and both factuality reranking methods (*linear* & *satu*) (y-axis). The horizontal line marks the baseline (points above are gains). Each point is a query.

MAP or BPREF. This could mean that, even after factuality reranking, some relevant documents are placed lower than the top ranks. For early precision, factuality reranking seems very effective, meaning that it pushes relevant documents even higher up in the ranks 1-10. This is important because, for the Web tracks, early precision measures are possibly more important than deep precision measures [28]. In addition, MAP is severely limited for the Web tracks [10] because it uses in its calculation a binary feature $rel(k)$ of value 1 if the kth ranked document is relevant, otherwise 0. However, $rel(k)$ is unknown for most $k > 12$ in ClueWeb09B, and methods to estimate MAP with incomplete knowledge of $rel(k)$ have proven to be unreliable [10].

The overall gains in precision, both averaged over the top 1000 documents, and in the single first relevant and top 10 retrieved documents, mean that factuality benefits ranking across the range of relevant documents in web search (those retrieved in the top ranks and those retrieved further down).

5.3.2   Per query analysis

Breaking down MAP and NDCG on a per query basis (Figure 5) illustrates the difference between curated and uncurated data even more clearly: for factuality reranking in curated data the gain is slightly lower than the deterioration (within -0.1 and +0.09 for MAP; within -0.2 and +0.17 for

NDCG). For uncurated data, the gain is always higher than deterioration, the largest being within -0.01 and +0.48 for MAP. It seems that in uncurated data, there may be more potential for improvement over the baseline with factuality. Figure 5 also shows that the precision gains reported in Table 3 are not misleadingly inflated by outliers that may affect the means of the evaluation measures, but rather they spread over most queries.

To look at which queries improve or deteriorate more closely, Table 4 groups queries by their % difference from the baseline, and shows the average MAP and NDCG of the baseline. I.e., the row '>10% .1310 (28)' means that there are 28 queries in Disks4-5 that improve by >10% over the baseline, and that the average baseline MAP of these queries is .1310. We see some interesting trends: queries with low baseline MAP (i.e. harder queries) improve by >10% for both curated and uncurated data; queries with higher baseline MAP (i.e. easier queries) improve by 0.1-10% for both curated and uncurated data; the queries which deteriorate are those whose baseline is closest to the mean for curated data, and those with the lowest MAP (i.e. the hardest) for uncurated data. These trends do not hold for NDCG and curated data, where queries of higher baseline scores either gain a bit (0.1-4.9%) or deteriorate; for uncurated data however we see the same trend as in MAP: only the hardest queries (of the lowest baseline score) deteriorate. The hardest queries are 22/150 ≈15% of all Web queries. Such extremely hard queries are often very vague or ambiguous, i.e. their difficulty is due more to their formulation, than to the quality of the documents retrieved for them. One plausible explanation is thus that factuality, which is a document quality, cannot compensate for such ill-formed queries.

5.3.3   Parameter sensitivity

We investigate how sensitive is the effectiveness of factuality reranking with *linear* and *satu* to perturbations of their respective parameters α and w. If slight parameter shifts cause effectiveness to drastically change, then the respective approach is sensitive, hence more prone to drastic changes in effectiveness if a poor parameter setting is chosen (even though sensitive methods do not necessarily have poor generalisation properties if properly tuned [46]). Figure 4 shows the geometry of the metric surface of the parameters used in *linear* and *satu* factuality reranking for MAP and NDCG for ClueWeb09B. The metric surface of w (*satu*) is relatively flat across the parameter space, hence more stable than the metric surface of α (*linear*), which has regions falling off to values of lower effectiveness. Similar plots are produced for factuality reranking in Disks4-5 (omitted for brevity). Hence, *satu*, not only performs better, but also seems to be more stable than *linear*. The stability of *satu* is also reported in [11].

6.   DISCUSSION

6.1   Correlation limitations

One should be careful when interpreting the correlations presented in Section 4 as a measure of association between factuality or objectivity and relevance. Correlation does not imply causation; we should not assume that a document is relevant because it is factual or objective. Nor should we assume the contrary, i.e. that there is no connection between factuality and objectivity, just because we find no

statistical correlation between them (Spearman's ρ: -0.03, P-value: 0.5 for Disks4-5; Spearman's ρ: 0.01, P-value: 0.8 for ClueWeb09B). Linguistic studies find clear links between factuality and objectivity [17, 50].

Furthermore, some aspects of bivariate relationships are not adequately addressed by correlation, e.g. changes in scale and location, because it is a symmetrical measure [16]. This means that effects attributable to location and differential variability, e.g. differences in importance between two variables, are not detected by correlation. Indeed, the scale differences between factuality/objectivity and relevance in Figure 2 above are quite large: the range of $p(relevance)$ (0.01-0.05) is consistently lower than the range of $p(factuality)$ and $p(objectivity)$ (0-1). This is because the documents assessed as relevant by TREC are both scarce and sparse (known for ClueWeb09B [10]), an artefact of the test collection design. In search engines where relevance assessments are simulated by features that are easier to extract in large numbers, e.g. user clicks, it is likely there may be a higher proportion of relevant documents among those assessed. Practically this may reduce the scale difference we see in Figure 2, resulting in clearer trends.

6.2 Efficiency

The computational cost of detecting factuality & objectivity for IR is not prohibitive, especially considering that it can be done off-line during indexing. For factuality detection, the most expensive step is fact extraction with ReVerb ($\sim$100ms per document on a Intel(R) Xeon(R) CPU with a 2.00GHz processor). For objectivity detection, the most expensive step is parsing with the StanfordParser (20s per document on average, on an Intel(R) Xeon(R) CPU with a 2.50GHz processor). For both factuality & objectivity, the exact computational complexity of LIBSVM is not known, but it is not linear [25]. These costs can be reduced considerably by parallelising the processing or improving the efficiency of the NLP code.

7. CONCLUSION

We presented a study of the relation between document factuality & objectivity and document relevance for IR. We used state-of-the-art factuality & objectivity detection from NLP on two TREC collections, one curated (official documentation/newswire) and one uncurated (web crawl). We found that (i) factuality is positively correlated to relevance, and the correlation is stronger for uncurated data; (ii) objectivity is not correlated to relevance for curated data, while, for uncurated data, the high presence of spam hinders its detection. We further experimented with factuality & objectivity as document scores, which we combined linearly with competitive baseline scores to rerank the top 1000 search results of 350 TREC queries. Factuality was found to be highly effective, especially for uncurated data, gaining >10% in average precision. Objectivity findings were mixed and inconclusive.

To our knowledge, this is the first study of document factuality & objectivity for back-end IR. Next steps include repeating the objectivity analysis on spam-free (or spam-light) uncurated data. We did not attempt this for ClueWeb09B because removing its spam would also remove ca. half of its TREC assessed documents, leaving us with very few documents to study ($\sim$42K or 0.08% of the collection). Another interesting future research direction is to explore how best to incorporate factuality and objectivity document scores to retrieval, e.g. by replacing the linear score combinations we used here for reranking with supervised learning for computing optimal rerankings, given the original ranking and the factuality or objectivity scores.

8. ACKNOWLEDGMENT

9. REFERENCES

[1] K. Al Khatib, H. Schütze, and C. Kantner. Automatic detection of point of view differences in Wikipedia. In *COLING*, pages 33–50, 2012.

[2] C. Banea, R. Mihalcea, and J. Wiebe. Sense-level subjectivity in a multilingual setting. *Computer Speech & Language*, 28(1):7–19, 2014.

[3] S. M. Beitzel, O. Frieder, E. C. Jensen, D. Grossman, A. Chowdhury, and N. Goharian. Disproving the fusion hypothesis. In *SAC*, pages 823–827, 2003.

[4] M. Bendersky, W. B. Croft, and Y. Diao. Quality-biased ranking of web documents. In *WSDM*, pages 95–104, 2011.

[5] J. Blumenstock. Word count as a measure of quality on wikipedia. In *WWW*, pages 1095–1096, 2008.

[6] B. Brost, I. J. Cox, Y. Seldin, and C. Lioma. An improved multileaving algorithm for online ranker evaluation. In *SIGIR*, pages 745–748, 2016.

[7] B. Brost, I. J. Cox, Y. Seldin, and C. Lioma. Multi-dueling bandits and their application to online ranker evaluation. In *CIKM*, page in press, 2016.

[8] A. Canestrelli, W. Mak, and T. Sanders. Causal connectives in discourse processing: How differences in subjectivity are reflected in eye movements. *Lang. & Cogn. Proc.*, 28(9):1394–1413, 2013.

[9] M. Coleman and T. L. Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, (60):282–284, 1975.

[10] G. V. Cormack, M. D. Smucker, and C. L. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *IRJ*, 14(5):441–465, 2011.

[11] N. Craswell, S. Robertson, H. Zaragoza, and M. Taylor. Relevance weighting for query independent evidence. In *SIGIR*, pages 416–423, 2005.

[12] R. Ennals, D. Byler, J. Agosta, and B. Rosario. What is disputed on the web. In *WICOW*, pages 67–74, 2010.

[13] O. Etzioni, M. Banko, S. Soderland, and D. Weld. Open information extraction from the web. *CACM*, 51(12):68–74, 2008.

[14] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *EMNLP*, pages 1535–1545, 2011.

[15] D. Fallows. Search engine users. *Pew Internet and American Life Project*, pages 1–29, 2005.

[16] R. A. Fisher. *Statistical Methods for Research Workers (13th ed.)*. Oliver and Boyd, London, 1958.

[17] T. Froehlich. *Information Ethics*. Encyclopedia of Library and Information Sciences. Routledge, 2002.

[18] M. Gamon, S. Basu, D. Belenko, D. Fisher, M. Hurst, and A. C. König. BLEWS: using blogs to provide context for news articles. In *ICWSM*, 2008.

[19] R. Gunning. *The technique of clear writing.* McGraw-Hill, 1952.

[20] E. Hargittai, L. Fullerton, E. Menchen-Trevino, and K. Thomas. Trust online: Young adults' evaluation of web content. *Int. J. of Com.*, 4(0):468–494, 2010.

[21] V. Hatzivassiloglou and K. McKeown. Predicting the semantic orientation of adjectives. In *ACL*, pages 174–181, 1997.

[22] L. Herzig, A. Nunes, and B. Snir. An annotation scheme for automated bias detection in wikipedia. In *Linguistic Annotation Workshop*, pages 47–955. 2011.

[23] F. Hill and A. Korhonen. Concreteness and subjectivity as dimensions of lexical meaning. In *ACL*, pages 725–731, 2014.

[24] C. Horn, A. Zhila, A. F. Gelbukh, R. Kern, and E. Lex. Using factual density to measure informativeness of web documents. In *NODALIDA*, pages 227–238, 2013.

[25] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. *LIBSVM: A Practical Guide to Support Vector Classification.* http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf, 2003.

[26] A. Juffinger, M. Granitzer, and E. Lex. Blog credibility ranking by exploiting verified content. In *WICOW*, pages 51–58, 2009.

[27] T. Kanungo and D. Orr. Predicting the readability of short web summaries. In *WSDM*, pages 202–211, 2009.

[28] D. Kelly, X. Fu, and C. Shah. Effects of position and number of relevant documents retrieved on users' evaluations of system performance. *TOIS*, 28(2):1–29, 2010.

[29] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas for navy enlisted personnel. Technical report, NTIS, 1975.

[30] A. Kittur, B. Suh, and E. H. Chi. Can you ever trust a wiki? In *CSCW*, pages 477–480, 2008.

[31] J. Kwok. Moderating the outputs of SVM classifiers. *Trans. on Neur. Netw.*, 10(5):1018–1031, 1999.

[32] R. Langacker. Deixis & subjectivity. In *The epistemic footing of deixis & reference*, pages 1–28, 2002.

[33] B. Larsen, C. Lioma, I. Frommholz, and H. Schütze. Preliminary study of technical terminology for the retrieval of scientific book metadata records. In *SIGIR*, pages 1131–1132, 2012.

[34] E. Lex, A. Juffinger, and M. Granitzer. Objectivity classification in online media. In *Hypertext and Hypermedia*, pages 293–294, 2010.

[35] E. Lex, M. Voelske, M. Errecalde, E. Ferretti, L. Cagnina, C. Horn, B. Stein, and M. Granitzer. Measuring the quality of web content using factual information. In *WebQuality*, pages 7–10, 2012.

[36] C. Lioma, A. Kothari, and H. Schütze. Sense discrimination for physics retrieval. In *SIGIR*, pages 1101–1102, 2011.

[37] C. Lioma, B. Larsen, and W. Lu. Rhetorical relations for information retrieval. In *SIGIR*, pages 931–940, 2012.

[38] C. Lioma and I. Ounis. Extending weighting models with a term quality measure. In *SPIRE*, pages 205–216, 2007.

[39] C. Lioma, J. G. Simonsen, B. Larsen, and N. D. Hansen. Non-compositional term dependence for information retrieval. In *SIGIR*, pages 595–604, 2015.

[40] C. Lioma, F. Tarissan, J. G. Simonsen, C. Petersen, and B. Larsen. Exploiting the bipartite structure of entity grids for document coherence and retrieval. In *ICTIR*, pages 11–20, 2016.

[41] C. Lioma and C. J. K. van Rijsbergen. Part of speech n-grams and information retrieval. *Revue française de linguistique appliquée*, XIII(1):9–11, 2008.

[42] N. Lipka and B. Stein. Identifying featured articles in wikipedia. In *WWW*, pages 1147–1148, 2010.

[43] G. McClure. Readability formulas: Useful or useless. *Trans. Prof. Comm.*, 30:12 – 15, 1987.

[44] G. H. McLaughlin. Smog grading – a new readability formula. *J. of Reading*, 12(8):639 – 646, 1969.

[45] M. J. Metzger. Making sense of credibility on the web. *JASIST*, 58(13):2078–2091, 2007.

[46] D. Metzler. Estimation, sensitivity, and generalization in parameterized retrieval models. In *CIKM*, pages 812–813, 2006.

[47] L. Michelbacher, A. Kothari, M. Forst, C. Lioma, and H. Schütze. A cascaded classification approach to semantic head recognition. In *EMNLP*, pages 793–803, 2011.

[48] M. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz. Understanding microblog credibility perceptions. In *CSCW*, pages 441–450, 2012.

[49] H.-J. Oh, Y.-C. Yoon, and H. K. Kim. Finding more trustworthy answers. *J. Inf. Sci.*, 39(4):509–522, 2013.

[50] C. E. Osgood, G. Suci, and P. Tannenbaum. *The Measurement of Meaning.* UIP, 1957.

[51] S. Park, S. Kang, S. Chung, and J. Song. NewsCube: Delivering multiple aspects of news to mitigate media bias. In *CHI*, pages 443–452, 2009.

[52] C. Petersen, C. Lioma, J. G. Simonsen, and B. Larsen. Entropy and graph based modelling of document coherence using discourse entities: An application to IR. In *ICTIR*, pages 191–200, 2015.

[53] C. Petersen, J. G. Simonsen, and C. Lioma. Power law distributions in information retrieval. *ACM TOIS*, 34(2):8, 2016.

[54] J. C. Platt. Probabilistic outputs for SVMs and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*, 1999.

[55] E. Riloff and R. Jones. Learning dictionaries for information extraction by multi-level bootstrapping. In *AAAI/IAAI*, pages 474–479, 1999.

[56] J. Schwarz and M. R. Morris. Augmenting web pages and search results to support credibility assessment. In *CHI*, pages 1245–1254, 2011.

[57] D. W. Scott. On optimal and data-based histograms. *Biometrika*, 66:605–610, 1979.

[58] J. G. Shanahan, Y. Qu, and J. Wiebe. *Computing Attitude and Affect in Text.* Springer, 2006.

[59] M. Thelen and E. Riloff. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *EMNLP*, pages 214–221, 2002.

[60] J. Wiebe and E. Riloff. Finding mutual benefit

between subjectivity analysis and information extraction. *T. Aff. Comp.*, 2(4):175–191, 2011.

[61] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Lang. Res. & Eval.*, 39(2-3):165–210, 2005.

[62] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. OpinionFinder: A system for subjectivity analysis. In *EMNLP*, pages 34–35, 2005.

[63] A. Yessenalina, Y. Yue, and C. Cardie. Multi-level structured models for document-level sentiment classification. In *EMNLP*, pages 1046–1056, 2010.

[64] Y. Zhou and W. B. Croft. Document quality models for web adhoc retrieval. In *CIKM*, pages 331–332, 2005.

[65] X. Zhu and S. Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In *SIGIR*, pages 288–295, 2000.