Linguistic Benchmarks of Online News Article Quality

Ioannis Arapakis[*]
Eurecat
Barcelona, Spain
arapakis.ioannis@gmail.com

Filipa Peleja[*]
Vodafone
Lisbon, Portugal
filipapeleja@gmail.com

B. Barla Cambazoglu[*]
Independent Researcher
barla@berkantbarlacambazoglu.com

Joao Magalhaes
NOVA-LINCS, DI, FCT
Universidade NOVA Lisboa, Portugal
jmag@fct.unl.pt

Abstract

Online news editors ask themselves the same question many times: *what is missing in this news article to go online?* This is not an easy question to be answered by computational linguistic methods. In this work, we address this important question and characterise the constituents of news article editorial quality. More specifically, we identify 14 aspects related to the content of news articles. Through a correlation analysis, we quantify their independence and relation to assessing an article's editorial quality. We also demonstrate that the identified aspects, when combined together, can be used effectively in quality control methods for online news.

1 Introduction

A recent study[1] found that online news is nowadays the main source of news for the population in the 18-29 age group (71%), and as popular as TV in the 30-39 age group (63%). The readers appetite for high-quality online news result in an offer of thousands of articles published every day in the whole of the Web. For instance, it is not uncommon to find the same facts reported by many different online news articles. However, only a few of them actually grab the attention of the readers. Journalists and editors follow standardised discourse rules and techniques aiming at engaging the reader in the article's narrative of article (Louis and Nenkova, 2013).

Analysing the discourse of such articles is central to properly assessing the quality of online news (van Dijk and Kintsch, 1983). Defining the variables that computational linguistics should quantify is a challenging task. Several questions arise from this exercise. For example, what does the quality refer to? What makes a new article perceived as high quality by the editors/users? What aspects of an article correlate better with its perceived quality? Can we predict the quality of an article using linguistic features extracted from its content? These are the kind of questions we address in this paper.

To this end, we propose a linguistic resource and assessment methodology to quantify the editorial quality of online news discourse. We argue that quality is too complex to be represented by a single number and should be instead decomposed into a set of simpler variables that capture the different linguistic and narrative aspects of online news. Thus, we depart from current literature and propose a multidimensional representation of quality. The first contribution of this paper is a taxonomy of 14 different content aspects that are associated with the editor-perceived quality of online news articles. The proposed 14 aspects are the result of an editorial study involving professional editors, journalists, and computational linguists.

The second contribution of this paper is an expert-annotated corpus of online news articles obtained from a major news portal. This corpus is curated by the editors and journalists who annotated the articles with respect to the 14 aspects and to the general editorial quality. To confirm the independence and relevance of the proposed aspects, we perform a correlation analysis on this ground-truth to determine the strength of the associations between different aspects and article editorial quality. Our analysis shows that the editor-perceived quality of an article exhibits a strong positive correlation with certain aspects, such as

[*] This work was done while the authors were at Yahoo Labs Barcelona.

[1] http://www.people-press.org/2013/08/08/amid-criticism-support-for-medias-watchdog-role-stands-out

fluency and *completeness*, while it is weakly correlated with other aspects like *subjectivity* and *polarity*.

As a baseline benchmark, we investigate the feasibility of predicting the quality aspects of an article using features extracted from the article only. Our findings indicate that article editorial quality prediction is a challenging task and that article quality can be predicted to a varying degree, depending on the feature space. The proposed aspects can be used to control the editorial quality with a Root Mean Squared Error (RMSE) of 0.398 on a 5-point Likert-scale.

The rest of the paper is organised as follows. Next, we discuss existing literature in discourse analysis and text quality metrics. In Section 3, we present the aspects that we identified as potential indicators of article quality. Section 4 provides the details of our online news corpus targeting the aspects of editorial quality control. The results of the correlation analysis conducted between the identified aspects and article quality are presented in Section 4. In Section 5, we present a baseline benchmark to automatically infer individual aspects and editorial quality from online news.

2 Related Work

A very recent work related to ours is (Gao et al., 2014), where the authors try to predict the interestingness of a news article for a user who is currently reading another news article. In our work, however, we try to predict the perceived quality of an article without using any context information other than the content of the article itself. Moreover, while the authors of (Gao et al., 2014) take a quite pragmatic approach to handle the problem, we follow a more principled approach and model the quality of a news article according to five orthogonal dimensions: readability, informativeness, style, topic, and sentiment. Work has been done in each one of these dimensions, but none has tackled the problem of modelling overall article quality in a comprehensive and articulated manner as we do. Below, we provide a survey of the previous work on these dimensions.

The readability of a piece of text can be defined as the ease that the text can be processed and understood by a human reader (Richards and Schmidt, 2013; Zamanian and Heydari, 2012). The readability is usually associated with fluency and writing quality (Nenkova et al., 2010; Pitler and Nenkova, 2008). Even though there is a significant amount of research that targets readability, most work (Redish, 2000; Yan et al., 2006) were originally designed to measure the readability of school books and do not suit well to more complex reading materials, such as news articles, which form the focus of our work.

The informativeness of a news article has been tackled from several different angles. In (Tang et al., 2003), news information quality was characterised by a set of nine aspects that were shown to have a good correlation with textual features. Catchy titles were shown to often lead to frustration, as the reader does not get the content that she expects (Louis and Nenkova, 2011). The task of assessing a news title's descriptiveness is related to semantic text similarity and has been researched by the SemEval initiative (Agirre et al., 2013). Moreover, the completeness of a news article is an aspect that has been considered in the past by (Louis and Nenkova, 2014), which showed that reporting the news with adequate detail is key to provide the reader with enough information to grasp the entire story. The freshness of news information also sets the tone of the discourse: information can be novel to the average reader or it can be already known and be presented as a reference to the reader. The novelty of an article is essentially accomplished by either analysing previous articles (Gamon, 2006) or by relying on real-time data from social-media services (Phelan et al., 2009).

The characterisation of the style of text compositions has been an active topic of research in communication sciences and humanities. An excellent example of the research done in this area is the influential work in (McNamara et al., 2009), where the authors found the best predictors of writing quality to be the syntactic complexity (number of words before the main verb), the diversity of words used by the author, and some other shallow features. In NLP, the writing style has been investigated in several contexts. A problem relevant to the one we addressed is the characterisation of an author's writing style to predict the success of novels (Ashok et al., 2013). The authors investigated a wide range of complex linguistic features, ranging from simple unigrams to distribution of word categories, grammar rules, distribution of constituents, sentiment, and connotation. The comparison of novels and news articles revealed a great similarity in the writing style of novels and informative articles.

The broadness of a news topic has an impact on the reader's perceived quality of the article. A technical article is usually targeting niche groups of users and a popular article targets the masses. One of the few corpus (Louis and Nenkova, 2013) addressing quality was limited to the domain of scientific journalism, thus more technical articles. This corpus only considered news from the New York Times, thus contained already very good quality news. Two recent work investigated the feasibility of predicting news articles' feature popularity in social media at cold start (Bandari et al., 2012; Arapakis et al., 2014a). In (Bandari et al., 2012), features extracted from the article's content as well as additional meta-data was used to predict the number of times an article will be shared in Twitter after it went online. In (Arapakis et al., 2014a), a similar study was repeated to predict the popularity of a news article in social media using additional features obtained from external sources.

Sentiment analysis concerns the subjectivity and the strength and sign of the opinions expressed in a given piece of text. In (Arapakis et al., 2014b), it was demonstrated that news articles exhibit considerable variation in terms of the sentimentality and polarity of their content. The work in (Phelan et al., 2009) has provided evidence that sentiment-related aspects are important to profile and assess the quality of news articles. Sentiment analysis has been applied to news articles in other contexts as well (Godbole et al., 2007; Balahur et al., 2010).

3 Modeling News Article Quality

The editorial control of news articles is an unsolved task that involves addressing a number of issues, such as identifying the characteristics of an effective text, determining what methods produce reliable and valid judgments for text quality, as well as selecting appropriate aspects of text evaluation that can be automated using machine learning methods. Underlying these tasks is a main theme: can we identify benchmarks for characterising news article quality? Therefore, there is a need for empirical work to identify the global and local textual features which will help us make an optimal evaluation of news articles.

By doing so, we achieve two goals. On one hand, we can offer valuable insights with respect to what constitutes an engaging, good quality news article. On the other hand, we can identify benchmarks for characterising news article quality in an automatic and scalable way and, thus, predict poor writing before a news article is even published. This can help reduce greatly the burden of manual evaluation which is currently performed by professional editors.

3.1 Methodology

The methodology described here provides a framework for characterising and modelling news article editorial quality. In our work, we follow a bottom-up approach and identify 14 different content aspects that are good predictors (as we demonstrate in Section 6.1) of news article quality. The aspects we identified are informed by input from news editors, journalists and computational linguists, and previous research in NLP and, particularly, the efforts in text summarisation (Bouayad-Agha et al., 2012), document understanding (Dang, 2005; Seki et al., 2006) and question answering (Surdeanu et al., 2008; Shtok et al., 2012).

After discussing the editorial quality control with professionals, we gathered a set of heuristics and examined the literature for ways of designing quantitative measures to achieve our goal. We group the aspects under five headings: readability, informativeness, style, topic, and sentiment (see Fig. 1). Below, we provide a brief description of each aspect.

3.2 Readability

High quality articles are written in a way that makes them easier to read. In our model, we include two different aspects related to readability (Pitler and Nenkova, 2008): fluency and conciseness.

Fluency: Fluent articles are built from sentence to sentence, forming a coherent body of information. Consecutive sentences are meaningfully connected. Similarly, paragraphs are written in a logical sequence.

Conciseness: Concise articles have a focus. Sentences contain information that is related to the main theme of the article. The same or similar information is tried to be not repeated.

3.3 Informativeness

As a main reason for reading online news is to remain well-informed (Tang et al., 2003), informativeness of articles have an effect on their per-

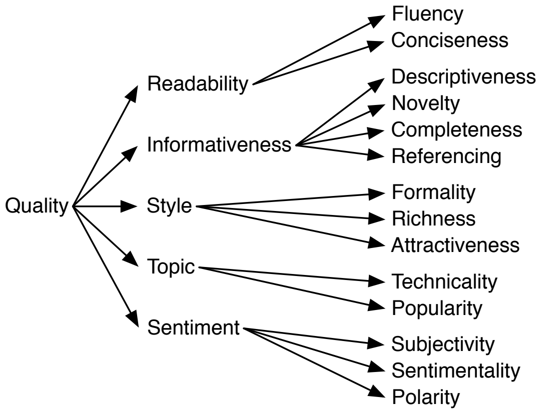

Figure 1: A taxonomy of the identified aspects.

ceived quality. In our model, we consider four different aspects related to informativeness: descriptiveness, novelty, completeness, and referencing.

Descriptiveness: Descriptiveness indicates how well the title of an article reflects its main body content. Titles with low descriptiveness are often click baits (e.g., "You won't believe what you will see"). Such titles may lead to dissatisfaction, as the provided news content usually does not meet the raised user expectation.

Novelty: Novel articles provide new and valuable information to the readers. The provided information is unlikely to be known to an average reader.

Completeness: Complete articles cover the topic in an adequate level of detail (Louis and Nenkova, 2014; Bouayad-Agha et al., 2012). A reader can satisfy her information need after reading such an article.

Referencing: Referencing is about the degree to which the article references external sources (including other people's opinions and related articles). Providing references allows the reader to access related information sources easily, (Gamon, 2006; van Dijk and Kintsch, 1983).

3.4 Style

The language and aesthetics is also related to the article quality (McNamara et al., 2009; Ashok et al., 2013; Pavlick and Tetreault, 2016; Peterson et al., 2011). We consider three style-related aspects: formality, richness, and attractiveness.

Formality: Formal articles are written by following certain writing guidelines. They are more likely to contain formal words and obey punctuation/grammar rules(Peterson et al., 2011).

Richness: The vocabulary of rich articles is perceived as diverse and interesting by the readers. Rich articles are not written in a plain and straightforward manner.

Attractiveness: Attractiveness measures the degree to which the title of an article raises curiosity in its readers. Attractive titles entice people to continue reading the main content of the article.

3.5 Topic

Editors consider the nature of the article with respect to its target audience, i.e., according to the target audience (technical or popular) the other aspects may play a different role. We investigate two topic-related aspects: technicality and popularity.

Technicality: Technical articles (Louis and Nenkova, 2013) usually require some effort to understand as well as previous knowledge on the topic. Examples of usually technical news topics include science and finance.

Popularity: The popularity refers to the size of the audience who would be interested in the topic of the article (Bandari et al., 2012; Arapakis et al., 2014b). For example, while many readers are interested in reading about celebrities, few readers are interested in articles about anthropology.

3.6 Sentiment

Finally, we consider the sentiments expressed in an article. Besides opinion articles (which are subjective by nature), many news may also convey a particular emotion. We evaluate three sentiment-related aspects: subjectivity, sentimentality, and polarity.

Subjectivity: Subjective articles tend to contain opinions, preferences, or possibilities. There are relatively few factual statements.

Sentimentality: Sentimentality is a measure of the total magnitude of positive or negative statements made in the article regarding an object or an event. Highly sentimental articles include relatively few neutral statements.

Polarity: Polarity indicates the overall sign of the sentiments expressed in the article (Arapakis et al., 2014a). Articles with positive (negative) polarity include relatively more statements with positive (negative) sentiment.

4 Corpus: Editorial Quality Control

Our goal is to identify proxies of news article quality that can be learned and predicted in an automatic and scalable manner. To identify these proxies, we rely on the domain knowledge and human intuition of expert judges, whom we employ in a rigorous, crowdsourcing-based evaluation for generating a ground-truth dataset. Through an editorial study we create an in-domain, annotated news corpus that allows us to learn predictive models which can estimate accurately the perceived quality of news articles.

4.1 Online News Articles

Our analysis was conducted on a dataset consisting of $13,319$ news articles taken from a major news portal[2]. We opted for a single news portal to be able to extract features that are consistent across all news articles. The dataset was constructed by crawling news articles over a period of two weeks. During the crawling period, we connected to the RSS news feed of the portal every 15 minutes and fetched newly published articles written in English. The content of the discovered articles was then downloaded from the portal.

Each article is identified by its unique URI and stored in a database, along with some meta-data, such as article's genre, its publication date, and its HTML content. We applied further filtering on the initial set of 13,319 news articles. The word count distribution of the articles followed a bimodal pattern, with the bulk of the articles located around a mean value of 447.5. Using this value as a reference point, we removed articles that contain less than 150 or more than 800 words. We then sampled a smaller set of articles such that each of the most frequent 15 genres have at least 65 articles in the sample. This left us with 1,043 new articles, out of which a randomly selected set of 561 articles were used in the editorial study.

The selected news articles were preprocessed before the editorial study. The preprocessing was performed in two steps. First, we removed the boilerplate of HTML pages and extracted the main body text of news articles, using Boilerpipe (Kohlschütter et al., 2010). Second, we segmented the body text into sentences and paragraphs. For sentence segmentation, we used the Stanford CoreNLP library, which includes a probabilistic parser (Klein and Manning, 2003; Mihalcea and Csomai, 2007). For each news article we generated a body- and sentence- level annotation form (see example in the supplementary notes).

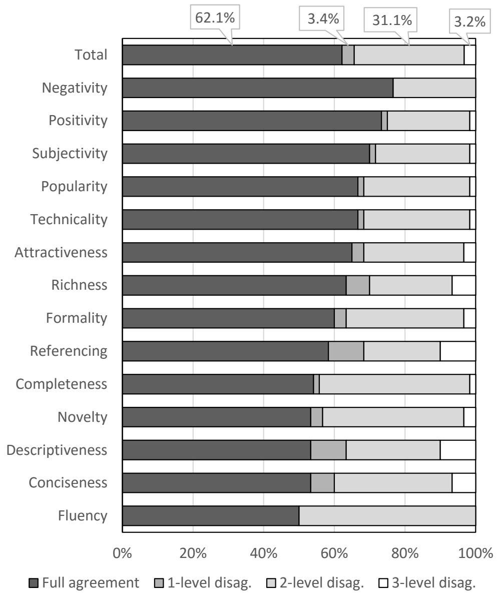

Figure 2: Annotators agreement.

4.2 Annotations of Editorial Quality Aspects

For our editorial study, we employed ten expert judges (male = 4, female = 6) who had a background in computational linguistics, journalism, or were media monitoring experts. The expert judges were either native English speakers or were proficient with the English language. The expert judges assessed a total of 561 news articles on 15 measures (14 aspects and the main quality measure), using a 5-point Likert scale, where low and high scores suggest weak or strong presence of the assessed measure, respectively.

The annotation took place remotely, and each expert judge could annotate up to ten news articles per day (this threshold was set to ensure a high quality of annotation), and each article was annotated by one expert judge and by one of the authors of this paper. Prior to that, there was a pilot session were each expert judge was asked to become familiar with the quality criteria and annotate three trial news articles. Next, a meeting (physical or online) was arranged and the authors discussed with the expert judge the rationale behind assigning the scores, and appropriate corrections and recommendations were made. This step ensured that we had disambiguated any questions prior to the editorial study and also assured that expert judges followed the same scoring procedure. The compensation for annotating was 10€per article. The annotated corpus is publicly available.[3]

Fig. 2 illustrates the details of the overall annotations agreement. We can see that annotations agree on 62.1% of the articles, on 65.5% they vary

[2] Yahoo News at http://www.yahoo.com/news.

[3] `http://novasearch.org/datasets/`.

only 1-point and in 96.6% they vary 2 points in the 5-point Likert-scale. These results are quite satisfying and show a good level of agreement and consistency across all aspects.

4.3 Corpus Statistics

Table 1 shows the mean (M) and standard deviation (SD) values for five different distributions (number of characters, words, unique words, entities, and sentences) and four different subsets of the corpus. The subsets contain all articles, high-quality articles (labels 4 and 5), medium-quality articles (label 3), or low-quality articles (labels 1 and 2). The last three subsets contain 84, 298, and 179 news articles, respectively. According to these numbers, the article quality follows an unbalanced distribution: about half of the articles are labeled as medium quality, and there are about two times more low-quality articles than high-quality articles. According to Table 1, there is a clear difference between distributions for the high- and low- quality articles. In general, we observe that higher-quality articles are relatively longer (e.g., more words or sentences), on average.

5 Aspects Correlation Analysis

To identify which aspects of a news article are better discriminants of its quality, we perform a correlation analysis. Given that we are looking at ordinal data that violates parametric assumptions, we compute the Spearman's rank correlation coefficients (r_s) between the aspects' scores and the news article quality that we acquired from our ground truth. The motivation behind this analysis is to get a first intuition into the aspects' effectiveness to act as quality predictors, by understanding how they are associated to news article quality.

In Table 2, we report several statistically significant correlations between the different aspects. Given that our correlation analysis involves multiple pairwise comparisons, we need to correct the level of significance for each test such that the overall Type I error rate (α) across all comparisons remains at .05. Given that the Bonferroni correction is too conservative in the Type I error rate, we opt for the more liberal criterion proposed by Benjamini and Hochberg (Benjamini and Hochberg, 1995; Benjamini and Hochberg, 2000) and compute the critical p-value for every pairwise comparisons as

$$p_{\text{crit}} = \frac{j}{k}\alpha, \tag{1}$$

where j is the index of all pairwise comparison p-values, listed in an ascending order, and k is the number of comparisons. If we consider Cohen's conventions for the interpretation of effect size, we observe that most of the correlation coefficients shown in Table 2 represent sizeable effects, which range from small ($\pm.1$) to large ($\pm.5$). For example, *completeness* is highly correlated with quality ($r_s = .70$) while *polarity* is the least correlated with quality ($r_s = .05$). In addition, Table 2 does not provide any evidence of multicollinearity since none of the aspects (with the exception of quality) are significantly highly correlated ($r_s > .80$).

6 Predicting Editorial Quality

6.1 Predicting EQ with the Aspects

In this section, we demonstrate the predictive characteristics of the proposed aspects (Section 3) with respect to news article quality. We formulate the prediction problem as a regression problem, and conduct a 10-fold cross validation to estimate the regression model. For our regression task we use a Generalised Linear Model (`GLM`) via penalized maximum likelihood (Friedman et al., 2010). The regularisation path is computed for the lasso or elasticnet penalty at a grid of values for the regularisation parameter lambda. The `GLM` solves the following problem

$$\underset{\beta_0,\beta}{min}\frac{1}{N}\sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda[(1-\alpha)\frac{\|\beta\|_2^2}{2} + \alpha\|\beta\|_1,], \tag{2}$$

over a grid of values of λ covering the entire range. Here $l(y, \eta)$ is the negative log-likelihood contribution for observation i. The elastic-net penalty is controlled by α, and bridges the gap between lasso ($\alpha = 1$, the default) and ridge ($\alpha = 0$). The tuning parameter λ controls the overall strength of the penalty. It is known that the ridge penalty shrinks the coefficients of correlated predictors towards each other while the lasso tends to pick one of them and discard the others, which makes it more robust against predictor collinearity and overfitting. We used the values that minimise RMSE, i.e., $\alpha = 0.95$ and $\lambda = 0.01$.

In Table 3, we see the coefficients of the final `GLM` model which are to be interpreted in the same manner as a Cox model. A positive regression coefficient for an explanatory variable means that the variable is associated with a higher risk of an event. In our case, all coefficients are positive, being *completeness*, *fluency* and *richness* the ones

Table 1: Statistics for the annotated news corpus (M ± SD values)

	All	High quality	Medium quality	Low quality
Characters	2490.28 ± 1900.95	4321.92 ± 2258.51	2698.94 ± 1641.85	1290.07 ± 1166.65
Words	413.03 ± 318.63	717.87 ± 386.08	447.46 ± 274.88	213.76 ± 193.12
Unique words	167.29 ± 110.25	269.08 ± 122.14	180.27 ± 95.72	98.29 ± 77.09
Entities	18.45 ± 14.43	23.85 ± 15.09	19.43 ± 11.03	14.29 ± 17.59
Sentences	20.67 ± 17.89	35.27 ± 24.92	21.76 ± 14.02	12.03 ± 14.42

Table 2: Correlations between different aspects in the ground-truth data

	Conc.	Desc.	Nov.	Comp.	Ref.	Form.	Rich.	Attr.	Tech.	Pop.	Subj.	Sent.	Pol.	Qual.
Fluency	$.61^{**}$	$.38^{**}$	$.34^{**}$	$.57^{**}$	$.37^{**}$	$.40^{**}$	$.53^{**}$	$.41^{**}$	$.15^{**}$	$.27^{**}$	$.12^{**}$	$.11^{**}$	$.03$	$\mathbf{.66}^{**}$
Conciseness		$.33^{**}$	$.32^{**}$	$.38^{**}$	$.28^{**}$	$.41^{**}$	$.39^{**}$	$.30^{**}$	$.24^{**}$	$.25^{**}$	$-.00$	$.08$	$.01$	$.47^{**}$
Descriptiveness			$.18^{**}$	$.32^{**}$	$.23^{**}$	$.23^{**}$	$.19^{**}$	$.13^{**}$	$.13^{**}$	$.17^{**}$	$.00^{*}$	$.09$	$.00$	$.37^{**}$
Novelty				$.39^{**}$	$.40^{**}$	$.44^{**}$	$.35^{**}$	$.37^{**}$	$.16^{**}$	$.32^{**}$	$.05$	$.25^{**}$	$-.05$	$.41^{**}$
Completeness					$.48^{**}$	$.38^{**}$	$.51^{**}$	$.39^{**}$	$.30^{**}$	$.26^{**}$	$.18^{**}$	$.20^{**}$	$.03$	$\mathbf{.70}^{**}$
References						$.51^{**}$	$.35^{**}$	$.33^{**}$	$.30^{**}$	$.29^{**}$	$.27^{**}$	$.44^{**}$	$.00$	$\mathbf{.52}^{**}$
Formality							$.43^{**}$	$.30^{**}$	$.46^{**}$	$.25^{**}$	$.01$	$.31^{**}$	$-.08$	$.47^{**}$
Richness								$.50^{**}$	$.25^{**}$	$.35^{**}$	$.24^{**}$	$.15^{**}$	$.04$	$\mathbf{.63}^{**}$
Attractiveness									$.15^{**}$	$.55^{**}$	$.28^{**}$	$.23^{**}$	$-.02$	$\mathbf{.52}^{**}$
Technicality										$.22^{**}$	$.16^{**}$	$.22^{**}$	$.11^{*}$	$.30^{**}$
Popularity											$.28^{**}$	$.24^{**}$	$.02$	$.41^{**}$
Subjectiveness												$.42^{**}$	$.11^{*}$	$.23^{**}$
Sentimemtality													$-.13^{*}$	$.27^{**}$
Polarity														$.05$

Significance levels (two-tailed) are as follows: * $:< .01$; ** $:< .001$.

Table 3: The coefficients of the final GLM model. The intercept value is 2.9103.

Group	**Aspects**	**Coefficients**
Readability	Fluency	.1730
	Conciseness	.0372
Informativeness	Completeness	.2062
	Descriptiveness	.0723
	Referencing	.0343
	Novelty	-
Style	Richness	.1192
	Formality	.0602
	Attractiveness	.0515
Topic	Popularity	.0578
	Technicality	.0047
Sentiment	Subjectivity	-
	Polarity	-
	Sentimentality	-

Table 4: Average performance across all ten folds for the GL model and for different feature sets.

Group	**Aspects**	**RMSE**	**RRSE**
All groups	All aspects	.3984	-
Readability	w/o Fluency	.4158	-4.36%
	w/o Conciseness	.3984	.00%
Informative.	w/o Completeness	.4233	-6.25%
	w/o Referencing	.4000	-.40%
	w/o Descriptiveness	.3999	-.37%
	w/o Novelty	.3981	-.07%
Style	w/o Richness	.4081	-2.43%
	w/o Attractiveness	.4009	-.62%
	w/o Formality	.3990	-.15%
Topic	w/o Popularity	.4003	-.47%
	w/o Technicality	.3976	.20%
Sentiment	w/o Subjectivity	.3974	.25%
	w/o Polarity	.3984	-.10%
	w/o Sentimentality	.3983	.02%

showing a higher relation to the overall editorial quality.

Next, we replicate our regression experiments for the GLM regression model, but this time we apply a leave-one-aspect-out method, to examine the relative importance of each aspect in explaining our predicted variable, i.e., the news article quality. To this end, we evaluate the 14 regression models, each one with out one of the aspects. The goal is to verify how prediction is affected by each individual quality aspect.

To compare the performance of our GLM regression model against the baseline method (with all quality aspects), we compute the Root Mean Squared Error (RMSE), given by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y} - y_i)^2}{N}} \qquad (3)$$

where $\hat{y}$ is the sample mean and y_i is the i-th estimate. However, while regression results give an idea of the prediction quality of the models they do not quantify the size of the difference of their performance. We, therefore, also compute the Root Relative Squared Error (RRSE) metric as it provides a good indication of any relative improvement over the baseline methods, given by

$$RRSE = 1 - \frac{RMSE_{GLM}}{RMSE_{Baseline}}. \quad (4)$$

Table 4 shows the RMSE and RRSE, with respect to the GLM regression model trained on all the features. These results show that *completeness*, *fluency* and *richness* are the aspects that most affect RMSE when they are missing from the full model.

6.2 Automatic Prediction of EQ

We examined a baseline model (BaselineM) that always predicts the mean value and a baseline GLM model (BaselineShallow) trained on shallow features, to automatically predict the editorial quality. Shallow or lexical features are commonly used in traditional readability metrics, which are based on the analysis of superficial text properties. Flesh-Kincaid Grade Level (Flesch, 1979; François and Fairon, 2012), SMOG (McLaughlin, 1969), and Gunning Fog (Gunning, 1952) are some examples of readability metrics. The simplicity of these features makes them an attractive solution compared to computationally more expensive features, such as syntactic (Feng et al., 2010). However, as Shriver (Schriver, 1989) points out, the readability metrics can be useful when used as gross index of readability. For our baseline, we consider the Flesh Kincaid, Coleman Liau, ARI, RIX, Gunning Fog, SMOG, LIX features.

In Table 5, we report the average performance of the GLM regression model, BaselineM, and BaselineShallow across all folds. We note that our GLM regression model improves the RMSE by at least 40%, compared to both baselines.

Finally, as a reference for future research with the proposed corpus, we trained GLM regression models to predict each aspect individually. Table 6 presents the RMSE for each aspect, for two different sets of feature: a standard BoW and the shallow features described previously, as well as the BaselineM. Despite the simplicity of the features, we can see that the aspects can be inferred from the articles. In particular, the model trained on the BoW features achieves an RMSE that is very close to that of the BaselineM, whereas the model trained on the shallow features outperforms all other models.

Table 5: Average performance across all ten folds for the GLM, BaselineM and BaselineShallow.

Method	RMSE	RRSE
BaselineM	0.7048	43.47%
BaselineShallow	0.8937	55.41%
GLM	0.3984	-

Table 6: Average performance across all ten folds for the GL model and for different feature sets.

Aspects	BoW	Shallow	BaselineM
Fluency	1.1571	1.1181	1.1462
Conciseness	1.2622	1.1968	1.2456
Completeness	.8408	.7945	.8130
Referencing	.7047	.6613	.7048
Descriptiveness	.9260	.8730	.9073
Novelty	.7994	.7607	.7797
Richness	.9866	.9454	.9568
Attractiveness	.7048	.6702	.6907
Formality	.7025	.6691	.6920
Popularity	.8329	.7825	.8250
Technicality	.7923	.7409	.7907
Subjectivity	.8750	.8283	.9094
Polarity	.8109	.7780	.8009
Sentimentality	.8170	.7668	.8046

7 Conclusions

In this paper, we proposed an annotated corpus for controlling the editorial quality of online news through 14 aspects related to editors perceived quality of news articles. To this end, we performed an editorial study with expert judges either in computational linguistics, journalism, or media monitoring experts. The judges assessed a total of 561 news articles with respect to 14 aspects. The study produced valuable insights. One important finding was that high quality articles share a significant amount of variability with several of the proposed aspects, which supports the claim that the proposed aspects may characterise news article quality in an automatic and scalable way. Another finding was that *fluency*, *completeness* and *richness* are the aspects that best correlate with quality, while *technicality*, *subjectivity* and *polarity* aspects show a poor correlation with quality. This shows that the text comprehension and writing style are aspects that are more relevant than sentiment. Later, we showed that using the entire set of 14 aspects we could predict the text quality with an RMSE of only 0.400 in a 5-point Likert-scale. This renders a very effective decomposition of news article quality into the 14 aspects. As future work, we plan to investigate other linguistic representations that can improve the automated extraction of the proposed aspects to better predict the article's perceived quality.

Acknowledgments

The first three authors were supported by MULTISENSOR project, partially funded by the European Commission, under the contract number FP7-610411. The fourth author was partially supported by FCT/MCTES through the project NOVA LINCS Ref. UID/CEC/04516/2013.

References

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-agirre, and Weiwei Guo. 2013. sem 2013 shared task: Semantic textual similarity, including a pilot on typed-similarity. In *In *SEM 2013: The Second Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics*.

Ioannis Arapakis, B.Barla Cambazoglu, and Mounia Lalmas. 2014a. On the feasibility of predicting news popularity at cold start. In LucaMaria Aiello and Daniel McFarland, editors, *Social Informatics*, volume 8851 of *Lecture Notes in Computer Science*, pages 290–299. Springer International Publishing.

Ioannis Arapakis, Mounia Lalmas, Berkant Barla Cambazoglu, Mari-Carmen Marcos, and Joemon M. Jose. 2014b. User engagement in online news: Under the scope of sentiment, interest, affect, and gaze. *JASIST*, 65(10):1988–2005.

Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. 2013. Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. 2010. Sentiment analysis in the news. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Valletta, Malta, may. European Language Resources Association (ELRA).

Roja Bandari, Sitaram Asur, and Bernardo A Huberman. 2012. The pulse of news in social media: Forecasting popularity. In *ICWSM*, pages 26–33.

Y. Benjamini and Y. Hochberg. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300.

Y. Benjamini and Y. Hochberg. 2000. On the adaptive control of the false discovery fate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25(1):60–83.

Nadjet Bouayad-Agha, Gerard Casamayor, Simon Mille, and Leo Wanner. 2012. Perspective-oriented generation of football match summaries: Old tasks, new challenges. *ACM Transactions on Speech and Language Processing (TSLP)*, 9(2):3.

Hoa Trang Dang. 2005. Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A Comparison of Features for Automatic Readability Assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING)*, COLING '10, pages 276–284, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rudolf Franz Flesch. 1979. *How to write plain English: A book for lawyers and consumers*. Harpercollins.

Thomas François and Cédrick Fairon. 2012. An AI readability formula for French as a foreign language. In *Proceedings of the Conference on Empirical Methods in Natural Language (EMNLP)*, pages 466–477. Association for Computational Linguistics.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Michael Gamon. 2006. Graph-based text representation for novelty detection. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, pages 17–24. Association for Computational Linguistics.

Jianfeng Gao, Patrick Pantel, Michael Gamon, Xiaodong He, Li Deng, and Yelong Shen. 2014. Modeling interestingness with deep neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena. 2007. Large-scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media*.

Robert Gunning. 1952. The Technique of Clear Writing. *McGraw-Hill*.

Dan Klein and Christopher D. Manning. 2003. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 441–450, New York, NY, USA. ACM.

Annie Louis and Ani Nenkova. 2011. Text specificity and impact on quality of news summaries. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 34–42. Association for Computational Linguistics.

Annie Louis and Ani Nenkova. 2013. A corpus of science journalism for analyzing writing quality. *Dialogue & Discourse*, 4(2):87–117.

Annie Louis and Ani Nenkova. 2014. Verbose, laconic or just right: A simple computational model of content appropriateness under length constraints. pages 636–644.

G Harry McLaughlin. 1969. SMOG grading: A new readability formula. *Journal of reading, JSTOR*, 12(8):639–646.

Danielle S McNamara, Scott A Crossley, and Philip M McCarthy. 2009. Linguistic features of writing quality. *Written Communication*.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 233–242, New York, NY, USA. ACM.

Ani Nenkova, Jieun Chae, Annie Louis, and Emily Pitler. 2010. Structural Features for Predicting the Linguistic Quality of Text. *Proceddings of the Empirical Methods in Natural Language Generation (EMNLP)*, 5790:222–241.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

Kelly Peterson, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: A case study on the enron corpus. In *Proceedings of the Workshop on Languages in Social Media*, pages 86–95. Association for Computational Linguistics.

Owen Phelan, Kevin McCarthy, and Barry Smyth. 2009. Using twitter to recommend real-time topical news. In *Proceedings of the third ACM conference on Recommender systems*, pages 385–388. ACM.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08.

Janice Redish. 2000. Readability formulas have even more limitations than Klare discusses. *ACM Journal of Computer Documentation (JCD)*, 24(3):132–137.

Jack C Richards and Richard W Schmidt. 2013. *Longman dictionary of language teaching and applied linguistics*, volume 78. Routledge.

Karen A Schriver. 1989. Evaluating text quality: The continuum from text-focused to reader-focused methods. *Professional Communication, IEEE Transactions on*, 32(4):238–255.

Yohei Seki, Koji Eguchi, Noriko Kando, and Masaki Aono. 2006. Opinion-focused summarization and its analysis at duc 2006. In *Proceedings of the Document Understanding Conference (DUC)*, pages 122–130.

Anna Shtok, Gideon Dror, Yoelle Maarek, and Idan Szpektor. 2012. Learning from the past: answering new questions with past answers. In *Proceedings of the 21st international conference on World Wide Web*, pages 759–768. ACM.

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2008. Learning to rank answers on large online qa collections. In *ACL*, volume 8, pages 719–727.

Rong Tang, Kwong Bor Ng, Tomek Strzalkowski, and Paul B Kantor. 2003. Automatically predicting information quality in news documents. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*, pages 97–99. Association for Computational Linguistics.

Teun A. van Dijk and W. Kintsch. 1983. *Strategies of discourse comprehension*. New York: Academic Press.

Xin Yan, Dawei Song, and Xue Li. 2006. Concept-based document readability in domain specific information retrieval. In *Proceedings of the 15th ACM Conference on Information and Knowledge Management (CIKM)*, pages 540–549. ACM.

Mostafa Zamanian and Pooneh Heydari. 2012. Readability of texts: State of the art. *Theory and Practice in Language Studies*, 2(1):43–53.