Assessing the Quality of Unstructured Data: An Initial Overview

Cornelia Kiefer

Graduate School of Excellence Advanced Manufacturing Engineering
Nobelstr. 12, 70569 Stuttgart, Germany
cornelia.kiefer@gsame.uni-stuttgart.de
http://www.gsame.uni-stuttgart.de

Abstract. In contrast to structured data, unstructured data such as texts, speech, videos and pictures do not come with a data model that enables a computer to use them directly. Nowadays, computers can interpret the knowledge encoded in unstructured data using methods from text analytics, image recognition and speech recognition. Therefore, unstructured data are used increasingly in decision-making processes. But although decisions are commonly based on unstructured data, data quality assessment methods for unstructured data are lacking. We consider data analysis pipelines built upon two types of data consumers, human consumers that usually come at the end of the pipeline and non-human / machine consumers (e.g., natural language processing modules such as part of speech tagger and named entity recognizer) that mainly work intermediate. We define data quality of unstructured data via (1) the similarity of the input data to the data expected by these consumers of unstructured data and via (2) the similarity of the input data to the data representing the real world. We deduce data quality dimensions from the elements in analytic pipelines for unstructured data and characterize them. Finally, we propose automatically measurable indicators for assessing the quality of unstructured text data and give hints towards an implementation.

Keywords: quality of unstructured data, quality of text data, data quality dimensions, data quality assessment, data quality metrics

1 Introduction

In recent years the methods for knowledge extraction from unstructured data have improved and unstructured data sources such as texts, speech, videos and pictures have gained importance. Nowadays, sentiment analysis of social media data leads to decisions in marketing campaign design, images are classified automatically and unstructured information can be retrieved easily using search engines [6, 19]. But methods which determine the quality of the data are lacking. To be able to make good decisions, the quality of the underlying data must be determined. Similar to the concepts, frameworks and systems developed for structured data we need means to ensure high quality of unstructured data. We focus on data consumers of unstructured data and define them as humans or non-humans / machines (e.g. algorithms) that are using or processing data. The quality of the data consumed by the final consumer such as a human who needs to derive a decision from the data, depends on the quality assessed for earlier consumers. This is especially true for unstructured data, which is analyzed in a pipeline.

The remainder of this paper is organized as follows: First we motivate research in assessing the quality of unstructured data in section 2. In section 3 we define data quality of unstructured data. Furthermore, we describe the data quality dimensions interpretability, relevance and accuracy. Based on this, in section 4 we present data quality indicators for unstructured text data. In section 5 we discuss related work and finally conclude the work and highlight future work in section 6.

2 Motivation

Low data quality is dangerous because it can lead to wrong or missing decisions, strategies and operations. It can slow down innovation processes, and losses for organizations caused by low data quality are estimated to lie over billions of dollars per year [8]. Bad data is a huge problem: 60% of enterprises suffer from data quality issues, 10-30% of data in organizational databases is inaccurate and individual reports of incomplete, inaccurate and ambiguous organizational data are numerous [13, 18].

The most important information sources in organizations, such as the workers, managers and customers produce unstructured data. About 90% of all data outside of organizations and still more than 50% inside are estimated to be unstructured [20]. In the era of Big Data the amount of data is increasing immensely and filtering relevant and high quality data gets more and more important. Organizations need to leverage the information hidden in unstructured data to stay competitive [14]. Therefore, the quality of texts, pictures, videos and speech data needs to be ensured. But while the need for data quality assessment and improvement strategies for unstructured data was recognized (e.g. [2, 23]) no concrete approach to assessing the quality of unstructured data was suggested yet. We fill this gap and provide data quality dimensions and executable indicators for unstructured data. By focusing on automatically calculable indicators of data quality, we aim to support real time analytics of stream data (such as social media data) with real time data quality assessment techniques, both running concurrently.

3 Definition of Data Quality and of Data Quality Dimensions for Unstructured Data

The definitions of data quality in [24, 30] focus on structured data which is consumed by humans. They define data quality via the similarity of the data

D to the data set D' which is expected by the data consumer [24] and via the fitness for use by the data consumer [30]. We extend the meaning of these existing definitions by pointing out that machine consumers and many different consumers in a pipeline need to be considered as well as human end consumers in the case of unstructured data. Furthermore, data quality needs to be defined in terms of accuracy. Accuracy describes the similarity between the input data and the data which would be representing the real world. This definition of Accuracy is equal to exiting ones, e.g. [11].

The quality of data has a multi-faceted nature and many lists of data quality dimensions and indicators for structured data exist (see 5). All of the dimensions that were found to be relevant in the literature, such as completeness, timeliness and accuracy are relevant to structured as well as unstructured data. From these dimensions we selected three dimensions which are relevant to mining processes on unstructured data.

We deduce the dimensions from the elements involved in mining processes on unstructured data: The input data, the real world, data consumers, a task and the knowledge extracted. Based on these elements, the quality of data D can be determined by comparing it to three classes of ideal data sets: the data as expected by the current data consumer D_C (we will call this the Interpretability dimension), the data as it would be optimal for the task D_T (Relevancy) and the data set which is representing the real world D_W (Accuracy). The deduced dimensions are also in line with the data quality definitions stated above. In Fig. 1, we illustrate the three data sets in the context of an ideal mining process on unstructured data. Ideally, D would match the real world D_W and would be exactly the same as the data expected by the first data consumer. Since unstructured data is analyzed in a pipeline, the output of the first data consumer is input to the second and should therefore match the data expected by the second data consumer and so on (as indicated in Fig. 1 with the analysis pipeline). An ideal result of the mining process can be D_T (which is still bound to D, D_W and D_C and is usually equal to the data expected by the final consumer). By basing the data quality dimensions on the elements involved in a mining process on unstructured data, we focus on the quality of unstructured data which is analyzed automatically in analytics pipelines.

In the following, we describe the deduced data quality dimensions in more detail:
Interpretability can be assessed as the degree of similarity between D and D_C. For example, consider a statistical preprocessor which is used to segment a text into sentences. If it was trained on Chinese texts and is used to segment English texts, D and D_C are not similar and data quality is low. Since often many different data consumers are involved in interpreting unstructured data, this dimension is crucial for unstructured data.

Relevancy can be assessed as the similarity between D and D_T. Usually D_T will be very similar to the D_C of the end consumer (which we will call D_{CE}) who wants to use the data to accomplish the task. While differences between D_T and the data expected by the end consumer D_{CE} indicate problems, these

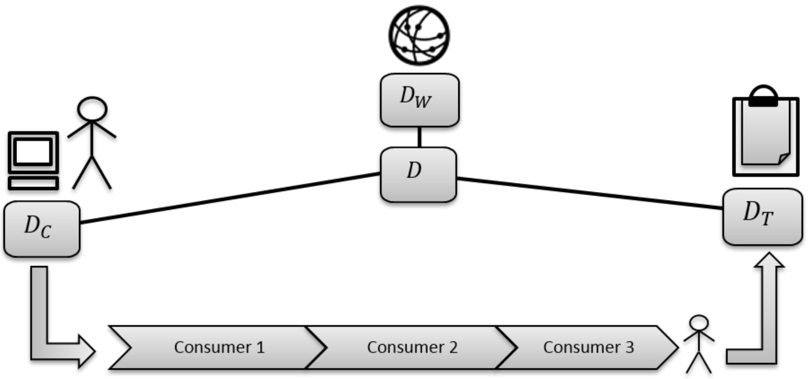

Fig. 1. The three ideal data sets D_C, D_T and D_W in the context of an ideal mining process on unstructured data

are not related to data quality and we will therefore assume D_T and D_{CE} to be equivalent. As an example for relevancy, consider a worker on the shop floor who is searching for a solution for an urgent problem with a machine in a knowledge base. If he only finds information on the price of the machine, the data quality of the result is low because it does not help him with his task of solving the problem.

We assess the **Interpretability and Relevancy** of a data set D by its similarity to the data set D_C and D_{CE} which is expected by the data consumers. Expectations differ from human to machine consumers. What a human data consumer expects, depends on factors such as his knowledge, experiences and goals. Expectations of machine consumers are very precise and depend on the algorithm, training data, statistical models, rules and knowledge resources available. This holds for all types of unstructured data. As illustrated in Fig. 2, unstructured data such as textual documents may be consumed by machines or humans and the data set D_C or D_{CE} depends on factors such as the native language of the human and the statistical language models available to the machine. For example, a human data consumer expects a manual for a machine to be in his native language or in a language he knows. He also expects the manual to explain the machine in a way he understands with his technical expertise. When a machine consumes unstructured data, similar factors influence the interpretability and more precisely the similarity of the input data and the data expected. The knowledge of a machine consumer can be represented by machine-readable domain knowledge encoded in semantic resources (such as taxonomies), by training data, statistical models or by rules. As an example, imagine a machine consumer that uses a simple rule-based approach to the extraction of proper names from German text data, where all uppercased words are extracted. This machine consumer expects a data set D_C with correct upper and lowercased words. If D is all lower-cased, D_C and D are not similar and the data is not fit for use by that data consumer.

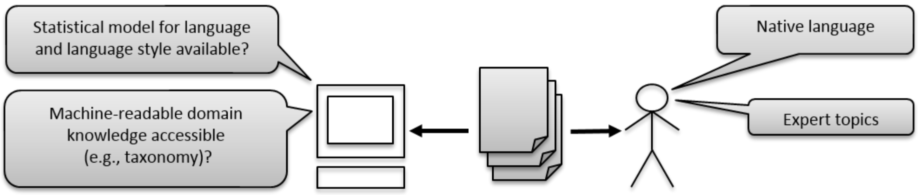
Statistical model for language and language style available?
Machine-readable domain knowledge accessible (e.g., taxonomy)?
Native language
Expert topics

Fig. 2. Machine and human data consumer and factors that influence the data expected

Unstructured data is usually consumed by many different data consumers with many different data sets D_C expected. In an analytics pipeline, the raw data is consumed and processed by several consumers in a row and the output of the previous consumer is the input to the next consumer and so on. Data quality problems at intermediate consumers may be automatically propagated to following consumers. By considering all intermediate (machine and/or human) consumers, the exact points for data quality improvement can be determined. In Fig. 3 we illustrate an analytics pipeline involving three machine consumers and one human end consumer of the data. Machine consumers are in this illustration represented by three high level machine consumers which are present in many analytic pipelines of unstructured data: preprocessors, classifiers and visualizers. For example, as depicted in Fig. 3, the output of the preprocessor is input to automatic classification and the results are then visualized. The visualizations are finally the input to a human consumer of the data, who e.g., derives decisions from it.

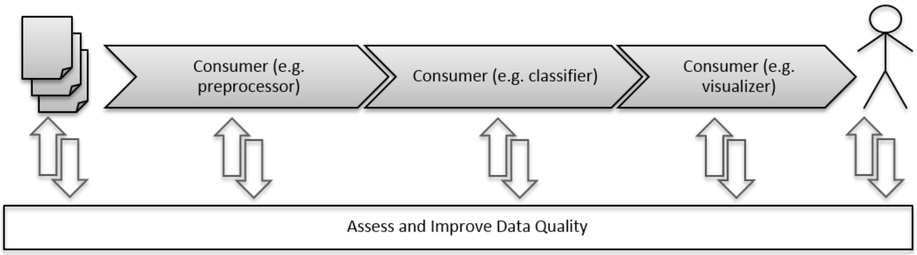
Consumer (e.g. preprocessor)
Consumer (e.g. classifier)
Consumer (e.g. visualizer)
Assess and Improve Data Quality

Fig. 3. Assessing and Improving data quality for each data consumer on the way from e.g., raw text documents to final consumer

As for structured data, the **Accuracy** of data and information is a very important data quality dimension. It is hard to measure, because the data set

D_W, which represents the real world, is often not known and creating it involves the work of human experts, is time-consuming, costly or even impossible. The solution is usually to abstract away from details e.g., by using rules to check general conformance of data points with expected patterns (e.g., e-mail addresses containing an @ sign) or to built D_W manually for a part of the data set only (see [28, 29]). D_W may be represented by a so-called gold standard data set with the accurate values annotated manually by human experts. For example, statistical classifiers are evaluated by comparing the prediction of the statistical classifier with those in a gold standard with manually annotated classes. Since D_W is not known for all data sets D, many statistical classifiers can not be evaluated and the number of problems with accuracy in big data bases can only be approximated.

4 Data Quality Indicators for Unstructured Text Data

A data quality dimension can be measured by exploitation of data quality indicators. Data quality indicators must be transferable to a number in the interval [0,1] where 0 indicates low data quality and 1 indicates high data quality (this is similar to the standard characterizations of data quality metrics, such as in [1]). Therefore, indicators can e.g., be represented by yes/no-questions, proportions of data items which have a certain characteristic or by evaluation metrics. The standard approaches to more concrete indicators for the quality of structured data involve counting the number of missing values, wrong values or the number of outliers. For the case of unstructured data, different indicators are needed. We compiled an extensive list of indicators for all three dimensions. The definition of indicators is based on the dimensions discussed in the previous section and on related work in natural language processing, information retrieval, automated assessment and machine learning (see section 5.2). Here, we limit the indicators presented to those which are (1) automatically measurable and (2) applicable to unstructured text data. Furthermore, we selected indicators, which we already implemented or which are straightforward to implement (since libraries with good documentations are available), so that the indicators can be verified in experiments in near future work. In table 1, we describe each dimension with these more concrete indicators of data quality.

While the concept behind the indicators *confidence*, *precision*, *accuracy* and *quality of gold annotations* are applicable to all types of unstructured data which are processed by statistical machine learning components, the remaining indicators are text specific. With a different definition of *noisy data* and *fit of training data*, the concepts may be transferred to other data types as well, e.g. measuring the similarity between input pictures and training data pictures or measuring the percentage of noisy data, defined as the percentage of background noise, in speech.

In the following we describe the indicators in more detail and give hints towards possible implementations:

Table 1. Indicators for the quality of unstructured text data

Dimension	Indicator
Interpretability	Fit of training data
	Confidence
	Noisy data
Relevancy	Frequent keywords
	Specificity
Accuracy	Precision
	Accuracy
	Quality of gold annotations

The first indicator *fit of training data* directly follows from the definition for **Interpretability** we gave in section 3, when considering statistical classifiers as data consumers. The quality of text data with respect to a machine consumer, can be measured by calculating the similarity of the input text data and the data expected by the data consumer. In the case of statistical classifiers such as a part of speech tagger (which automatically assigns parts of speech to each token such as a word in a text) or sentiment classifier (which automatically detects opinions in texts and assigns e.g., the classes positive, negative and neutral to texts), D_C may be represented by the training data. For the case of unstructured text data the similarity can be measured using text similarity measures. For example, consider the situation where Twitter data is consumed by a statistical classifier such as a part of speech tagger that was trained on newspaper texts. By the definition of interpretability used in this work, data quality is lower than for another tagger that was trained on text data from Twitter as well. Examples for measures for this indicator are text similarity measures such as Cosine Similarity and Greedy String Tiling which are e.g. implemented in the DKPro Similarity package (see [7]). Using the DKPro Similarity library in Java two lists of tokens can be easily compared and a similarity score in the interval [0,1] can be calculated, following the instructions on the web site[1].

The second indicator, *confidence*, also focuses on data quality of text data as perceived from the point of view of a statistical classifier. A statistical classifier estimates the probabilities for each class from a fixed list of classes, given the data. These probabilities are also called confidence values (for more details, see [12]). If the probability of a classification decision is very high, confidence of the statistical classifier is said to be high. Confidence is expressed as a number in the interval [0,1] and may be used for measuring data quality. For example, confidence measures are available and can be retrieved for the natural language processing tools in OpenNLP[2] (such as the tokenizer and part of speech tagger), a Java library for natural language processing which is heavily used in industry applications because it has an Apache license. To get these confidence values, follow the documentation of the OpenNLP library (see footnote 2, e.g., for the

[1] https://dkpro.github.io/dkpro-similarity/

[2] https://opennlp.apache.org/

part of speech tagger, just call the *probs* method which will return an array of the probabilities for all tagging decisions).

The third indicator in the interpretability dimension is the percentage of *noisy data*. This is a relevant indicator for human and machine consumers, since reading a text is more difficult for a human if it is full of misspelled words, non-grammatical sentences and abbreviations. Since most machine consumers of text data expect clean text data such as newspaper texts, the degree of noisy data also measures data quality from the viewpoint of such standard machine consumers. The percentage of noisy data may be measured as the percentage of sentences which cannot be parsed by an automatic syntax parser, unknown words, punctuation, very long/short sentences, incorrect casing, special signs, urls, mail addresses, emoticons, abbreviations, pause filling words, rare words or by the percentage of spelling mistakes (the latter as already suggested by [26]). Non-parsable sentences can be identified using an automatic syntax parser such as the parser implemented in natural language processing libraries such as OpenNLP (see footnote 2) or the Natural Language Processing Tool Kit NLTK[3]. The number of punctuation and of unknown words (e.g., defined as words unknown to a standard part of speech tagger) may be e.g., calculated using the standard part of speech tagger implemented in NLTK (which has individual classes for punctuation and unknown words). Very long/short sentences can be identified using a tokenizer and a sentence segmenter from a natural language processing library and by counting the automatically determined tokens and sentences. Incorrect casing may be detected using supervised machine learning methods, such as suggested in [17]. Regular expressions can be used to automatically identify the percentage of special signs, urls, mail adresses, emoticons, abbreviations and pause filling words in texts. Rare words can be identified internally by counting all words that occur less than a specified number of times in the text corpus, by counting words that are not found in a standard dictionary or a generated dictionary (such as a dictionary generated from a very encompassing text corpus from the domain). The number of spelling mistakes in a text corpus may be calculated using the Python implementation PyEnchant[4] or any other spelling correction module. Most of the measures suggested for the indicator noisy data can be implemented using the NLTK library which comes with very good documentation and an active community (see footnote 3).

But it is not sufficient if data is interpretable only. Interpretable data, which is not relevant to the end data consumer and his goal is of low quality. Therefore, it's **Relevancy** need to be calculated. For text data this can be done following approaches already developed for information retrieval systems. The relevance metric used in information retrieval systems determines the relevance of search results with respect to the information need of the searcher. The information need is captured via keywords or documents first and can then be compared e.g., to the *frequent keywords* in the input texts (see [16] for the relevance metric in information retrieval). Again, textual similarity measures such as cosine

[3] http://www.nltk.org/

[4] http://pythonhosted.org/pyenchant/

similarity are used to determine the similiarity of the information need and a text (as implemented in [7] and accessible via the well-documented DKPro Similarity library, see footnote 4). Besides the *frequent keywords*, also specificity can indicate the relevance of unstructured text data for the task a certain end consumer wants to accomplish. The *specificity* of language in texts and speech can be determined via the coverage of a domain-specific semantic resource which contains all relevant technical terms. In the simplest version this would be a text file with all domain words listed which is used to determine the percentage of domain words in a corpus. Coverage of domain specific taxonomies may be e.g., calculated with a concept matcher such as the one presented in [22].

If the data is interpretable and relevant, the remaining question is whether it reflects the real world or not, that is whether it is accurate. The **Accuracy** of unstructured text data may be indicated by evaluation metrics such as precision and accuracy. These metrics compare the automatically annotated data to parts of the data which represent the real world, such as manually annotated gold standard corpora. Statistical classifiers are evaluated by comparing them to gold standards and by determining how many of the classified entities really belong to a class (*precision*) and the percentage of classification decisions that were correct (*accuracy*), see [16]. The metrics precision and accuracy were already suggested as indicators for text data quality by [26] and [23]. Furthermore, the *quality of gold annotations* of training and test data is an indicator in the accuracy dimension. These can be calculated according to [10] by measuring the inter-rater agreement which measures the number of times one or more annotators agree. Evaluation metrics and inter-rater metrics are e.g. implemented in NLTK (see footnote 3).

In this section we presented automatically measurable indicators for text data which are executable. Not all indicators presented here are relevant and applicable in all cases. Only few out of the many statistical tools give access to the confidence metric and only with access to gold test data precision and accuracy can be calculated.

5 Related Work

While research on the quality of structured data is numerous, the quality of unstructured data has hardly been considered yet. We present related work in the field of data quality in section 5.1 and list isolated methods useful in assessing unstructured text data quality in section 5.2.

5.1 Related Work in Data Quality

Many frameworks and data quality dimensions dedicated to the quality of structured data have been suggested (e.g. [24, 30]) and also special frameworks and dimensions for social media data and big data were developed [5, 21]. In these works, data quality dimensions are defined from a human end consumer's point of view and no automatic measures for the assessment of unstructured data are given. Several sources [2, 23, 26] address the need for data quality measures on unstructured data but none of them gives executable dimensions and indicators. In these works, interesting starting points for quality dimensions and indicators are defined, such as:

- The quality of technologies used to interpret unstructured data and the author's expertise [23]
- Accuracy, readability, consistency and accessibility [2]
- Precision and spelling quality [26]

No hints towards possible implementations of these dimensions and indicators are suggested, though. As demanded in [26], we also support the view that textual data quality needs to be measured for both, human consumers and machine consumers. We have furthermore motivated the need to measure data quality at every stage. This is also demanded in [15, 27]. A closely related idea is also expressed in the concept of data provenance which aims at collecting the information on all data sources and transformation or merging steps of data (see [4]).

5.2 Isolated Methods for Data Quality Assessment of Unstructured Text Data

In the definition of the quality indicators in this article we focused on unstructured text data. Therefore, we limit the list of isolated methods to those relevant for the assessment of textual data. For example, quite some work in the field of natural language processing focuses on the interaction between textual data characteristics and the performance of Natural Language Processing (NLP) tools. In [3] the authors consider factors that affect the accuracy of automatic text-based language identification (such as the size of the text fragment and the amount of training data). Furthermore, work on correcting upper and lowercasing of words in texts (re-casing), spelling correction, abbreviation expansion and text simplification is related to our work (e.g., [17]). In the context of search engines, the quality of the search results and of the data basis is discussed as well [9]. In automated assessment, methods to automatically assess the quality of hand-written essays and short answers (e.g., student essays and answers to free text questions) are developed (for a good overview, see [31]). Work on training data selection in machine learning, which is on choosing subsets of training data which fit best to the domain of the test set (e.g. [25]) is also related to our work. The idea expressed in these works is similar to the idea behind the indicator *fit of training data*, which we added to our list of indicators for unstructured text data quality. However, we are the first to suggest the fit of training data as a data quality indicator. Furthermore, we do not suggest to use it for parts of training data, as suggested in these works, but to choose from different text corpora.

6 Conclusion and Future Work

We listed dimensions and indicators for determining the quality of unstructured data based on the basic elements of mining processes on unstructured data.

The indicators proposed are executable and easily transfer into a data quality metric in the interval [0,1]. In future work we will determine the most suitable implementations for the indicators and validate them in experiments. We will furthermore explore how indicators may be combined to measure the overall data quality of unstructured data and how the improvement of data quality as perceived by intermediate consumers influences data quality from a rather end consumer viewpoint.

Acknowledgments. The authors would like to thank the German Research Foundation (DFG) for financial support of this project as part of the Graduate School of Excellence advanced Manufacturing Engineering (GSaME) at the University of Stuttgart. Moreover, we thank B. Mitschang and L. Kassner for important feedback.

References

1. C. Batini, D. Barone, F. Cabitza, and S. Grega. A data quality methodology for heterogeneous data. *International Journal of Database Management Systems (IJDMS)*, 3(1):60–79, 2011.
2. C. Batini and M. Scannapieco. *Data and Information Quality*. Springer International Publishing, Cham, 2016.
3. G. R. Botha and E. Barnard. Factors that affect the accuracy of text-based language identification. *Computer Speech & Language*, 26(5):307–320, 2012.
4. P. Buneman and S. B. Davidson. Data provenance – the foundation of data quality. 2010.
5. L. Cai and Y. Zhu. The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14(0):2, 2015.
6. F. Camastra and A. Vinciarelli. *Machine learning for audio, image and video analysis: Theory and applications*. Advanced Information and Knowledge Processing. Springer, London, second edition edition, 2015.
7. Daniel Bär, Torsten Zesch, and Iryna Gurevych. Dkpro similarity: An open source framework for text similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations) (ACL 2013)*, pages 121–126, Stroudsburg, PA, USA, 2013. Association for Computational Linguistics.
8. D. Dey and S. Kumar. Reassessing data quality for information products. *Management Science*, 56(12):2316–2322, 2010.
9. C. Feilmayr. Decision guidance for optimizing web data quality - a recommendation model for completing information extraction results. *24th International Workshop on Database and Expert Systems Applications*, pages 113–117, 2013.
10. Fleiss and Levin. The measurement of interrater agreement. In J. L. Fleiss, B. Levin, and M. C. Paik, editors, *Statistical methods for rates and proportions*, Wiley series in probability and statistics, pages 598–626. J. Wiley, Hoboken, N.J., 2003.
11. C. Fox, A. Levitin, and T. Redman. The notion of data and its quality dimensions. *Inf. Process. Manage.*, 30(1):9–19, 1994.

12. S. Gandrabur, G. Foster, and G. Lapalme. Confidence estimation for nlp applications. *ACM Transactions on Speech and Language Processing (TSLP)*, 3(3):1–29, 2006.
13. J. Han, K. Chen, and J. Wang. Web article quality ranking based on web community knowledge. *Computing*, 97(5):509–537, 2015.
14. K. Hartl and O. Jacob. Determing the business value of business intelligence with data mining methods. *The Fourth International Conference on Data Analytics*, pages 87–91, 2015.
15. A. Immonen, P. Paakkonen, and E. Ovaska. Evaluating the quality of social media data in big data architecture. *IEEE Access*, (3):1, 2015.
16. C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval.* Cambridge University Press, New York, 2008.
17. C. Niu, W. Li, J. Ding, and R. K. Srihari. Orthographic case restoration using supervised learning without manual annotation. *International Journal on Artificial Intelligence Tools*, (13), 2003.
18. J. R. Nurse, S. S. Rahman, S. Creese, M. Goldsmith, and K. Lamberts. Information quality and trustworthiness: A topical state-of-the-art review. *International Conference on Computer Applications and Network Security (ICCANS 2011)*, 2011.
19. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86, 2002.
20. P. Russom. Bi search and text analytics: New additions to the bi technology stack. 2007.
21. M. Schaal, B. Smyth, R. M. Mueller, and R. MacLean. Information quality dimensions for the social web. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, pages 53–58. ACM, 2012.
22. M. Schierle and D. Trabold. Multilingual knowledge-based concept recognition in textual data. In A. Fink, B. Lausen, W. Seidel, and A. Ultsch, editors, *Advances in Data Analysis, Data Handling and Business Intelligence*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 327–336. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.
23. A. Schmidt, C. Ireland, E. Gonzales, M. Del Pilar Angeles, and D. D. Burdescu. On the quality of non-structured data, 2012.
24. L. Sebastian-Coleman. *Measuring data quality for ongoing improvement: A data quality assessment framework.* Elsevier Science, Burlington, 2013.
25. Y. Song, P. Klassen, F. Xia, and C. Kit. Entropy-based training data selection for domain adaptation. *Proceedings of COLING 2012*, 2012.
26. D. Sonntag. Assessing the quality of natural language text data. In *GI Jahrestagung*, pages 259–263, 2004.
27. I.-G. Todoran, L. Lecornu, A. Khenchaf, and J.-M. Le Caillec. A methodology to evaluate important dimensions of information quality in systems. *Journal of Data and Information Quality*, 6(2-3):1–23, 2015.
28. T. Vogel, A. Heise, U. Draisbach, D. Lange, and F. Naumann. Reach for gold. *Journal of Data and Information Quality*, 5(1-2):1–25, 2014.
29. H. Wang, M. Li, Y. Bu, J. Li, H. Gao, and J. Zhang. Cleanix. *ACM SIGMOD Record*, 44(4):35–40, 2016.
30. R. Y. Wang and D. M. Strong. Beyond accuracy: what data quality means to data consumers. *J. Manage. Inf. Syst.*, 12(4):5–33, 1996.
31. R. Ziai, N. Ott, and D. Meurers. Short answer assessment: Establishing links between research strands. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7)*, Montreal, Canada, 2012. Association for Computational Linguistics.