Fuzzy Based Text Quality Assessment for Sentiment Analysis

Manel BenSassi[1(✉)], Maher Abbes[1(✉)], and Faten Atigui[2(✉)]

[1] Univ. Manouba, ENSI, RIADI LR99ES26, Campus universitaire, 2010 Manouba, Tunisia
{manel.bensassi,maher.abbes}@ensi.uma.tn
[2] CEDRIC, Conservatoire National des Arts et des Métiers (CNAM) PARIS, Rue Saint Martin, 75003 Paris, France
faten.atigui@cnam.fr

Abstract. Practitioners have emphasized the importance of employing sentiment analysis techniques in decision-making. The data utilized in this process is typically gathered from social media, making it somewhat unreliable for decision-making. To address this issue, this study focuses on the Text Quality (TQ) aspect to capture the characteristics of Twitter data streams. Our objective is to develop an automated approach that assists the user in assessing the quality of textual data. This is accomplished through a fuzzified classifier, which automatically identifies ambiguous and unambiguous text at both the syntactic and semantic levels. We present a software tool that captures real-time and batch Twitter data streams. This tool calculates their TQ and presents the outcomes through diverse graphical depictions. It also empowers users to customize the weights allocated to individual quality dimensions and metrics used in computing the overall data quality of a tweet. This flexibility enables customization of weights according to different analysis contexts and user profiles. To demonstrate the usability and value of our contributions, we conducted a case study focusing on the Covid-19 vaccine. A preliminary analysis shows that by removing ambiguous text, the accuracy of the deployed algorithms enhances.

Keywords: Sentiment Analysis · Data Analytics · Data Quality · Big Data · Fuzzy Logic

1 Introduction

The literature provides evidence that furnishing decision-makers with trustworthy information has a positive impact on both tactical and strategic decisions. The growing need to discover and integrate reliable information from heterogeneous data sources, distributed in the Web, Social Networks, Cloud platforms or Data Lakes, makes Data Quality (DQ) an imperative topic. Becoming one of the most important elements in the decision-making process, sentiment analysis is concerned with gathering, analyzing, specifying and predicting user opinions

F. Zhang et al. (Eds.): WISE 2023, LNCS 14306, pp. 13–23, 2023.
https://doi.org/10.1007/978-981-99-7254-8_2

that are described in natural language for the most part. According to [1], there is a prevailing belief that the quality of social media data streams is commonly low and uncertain, which, to a certain extent, renders them unreliable for making decisions based on such data. Thus, to be used in decision-making scenarios, tweets should have a minimum quality to avoid deficient decisions. The main problem in extracting opinions from social media texts is that words in natural language are highly ambiguous.

Research Hypothesis. Our hypothesis is that errors introduced into sentiment analysis (and the consequent confidence decrease in decision making process based on sentiment analysis) are primarily attributed to the ambiguity present in the text. In our work, we use the term "ambiguity" in its more general sense: 1) The first aspect is "the capability of being understood in two or more possible senses or ways" [2] that derived from linguistic features such as poorly constructed sentences or syntactical errors [3] and, 2) "Uncertainty" [3] refers to the lack of semantic information and grounding between the writer and reader. Thus, with reference to the investigation done by [4] ambiguity could be classified into "syntactic" and "semantic" metrics. For this, our main research questions are the following:

- *How can we assess the TQ of streamed tweets in real and in batch time ?*
- *What are the relevant metrics and indicators to measure in order to ensure TQ?*

The aim of our research is to provide automated assistance for assessing the quality of textual data. To be used for different goals in different situations, context had to be given to data quality which means that data quality dimensions and metrics should be addressed differently in each case. Besides, we think that domain experts should be involved in the analysis process. Thus, it gives more flexibility to reuse our proposal in different contexts.

The research reported in this paper targets an automatic assessment of sentiment analysis text by means of a fuzzified classifier to automatically flag ambiguous and unambiguous text at syntactic and semantic level. Our approach considers textual data and consists of: (i) involving domain expert for a contextual analysis by allowing to change the weight of quality dimension metrics, (ii) evaluating tweets using text quality metrics especially ambiguity ones at real and batch time,(iii) and storing searches in a document-oriented database in order to ensure efficient information retrieval.

This paper is structured as follows. Section 2 gives an overview of sentiment analysis, and text quality related work. Section 3 presents our contribution for text quality dimensions and metrics. We present the experimental study in Sect. 4 before concluding in Sect. 5.

2 Related Work

Many issues have been highlighted, in the field of DQ and TQ in machine learning applications, such as the noisy nature of input data extracted from social

media sites [5] or insufficiency [6]. Other research on mining tweets regarding the irrelevance of data [5] and on performing sentiment analysis to discover the user's feeling behind a tweet, have been done in crisis times [7].

A more comprehensive analysis from DQ point of view, [1] proposed a DQ evaluation system based on computing only higher DQ dimensions and metrics for data streamed in real time. A DQ approach based on three strategies for social media data retrieval by monitoring the crawling process, the profiling of social media users, and the involvement of domain experts in the analytical process is advanced by [8]. [9] enhanced TQ through data cleansing model for short text topic modeling. However, most of the previous studies advance the DQ assessment as a crisp process based on quantitative data or statistical function which can reinforce difficulties for interpreting quality measure.

Other studies have considered that textual data couldn't be processed as certain input data [10]. For this, to handle uncertain and imprecise data, fuzzy ontology to assess the quality of linked data [11] and fuzzy knowledge-based system that combines the domain knowledge of an expert with existing measurement metric [12] were advanced.

Nevertheless, these approaches do not dive into rudimentary DQ dimensions and metrics and are closely tied to their context making their reuse heavy. We think that the challenges of TQ assessment remain into proposing an automatic evaluation approach having these main features: (i) adaptable and reusable according to the context of deployment through expert's involvement, (ii) extensible allowing the mashup of multiple fuzzy data sources, (iii) visualizing results at real and batch time, (iv) and based on hierarchical definition of multi-level quality dimensions and metrics explained in the following section.

3 Fuzzy Based Text Quality Assessment for Sentiment Analysis Approach

This section introduces our innovative automatic assessment approach that relies on a fuzzy tree classifier explained in Sect. 3.2 and a hierarchical definition of TQ dimensions and metrics introduced in Sect. 3.1.

3.1 Underlying Quality Model

Based on the proposed hierarchical definition of quality and its indicators in [13], we suggest enriching data quality metrics definition with text ambiguity metrics and context management as shown in Fig. 1. When dealing with text quality assessment, three main levels could be identified: `word`, `sentence`, and `discourse level`. Quality evaluation needs to be spread over these abstraction levels and consider the decision-making context. Besides, the hierarchical decomposition of the ambiguity concept into quantifiable indicators affecting the quality of the text could be adapted and adjusted according to different viewpoints.

For this purpose, we had to identify the discriminating features of the text that characterize the quality of social network text from syntactic and semantic point of view. We propose in Table 1, a formal definition of adopted syntactic and semantic ambiguity metrics. These metrics should be weighted by domain experts. We think that this proposal would provide : (i) flexibility, since domain experts can adapt to context variations, (ii) generality, since they can include many particular context-dependent cases, and (iii) richness, leading to include more aspects to the metric.

3.2 Fuzzy Tree-Based Classifier for Text Quality Assessment

To be used for different goals in different situations, data quality dimensions and metrics should be addressed differently in each case. For this, domain experts should be involved in the analysis process. Thus, it gives more flexibility to reuse our proposal in different contexts. So, our TQ assessment approach is based on the computing of TQ weighted metrics regulated by activation factors considering the context of deployment.

The TQ assessment, as depicted in Fig. 2, involves a two-phase process. The first pre-processing phase is elementary to establish necessary data for the quality computing phase. Hence, the pre-processing phase aims to set up (1) the weighted and activation set for TQ parameters and, (2) the conflict resolution when more than one expert are involving in the analytical process.

Based on those pre-established configuration and parameters, the quality assessment phase is divided into two main steps which are:(1) the computing of fuzzy metric and, (2) the inference of fuzzy decision tree, detailed as follow.

3.2.1 Pre-processing Phase

This phase is elementary to establish necessary data and parameters for the run-time execution of the system. In this section, we present our approach for

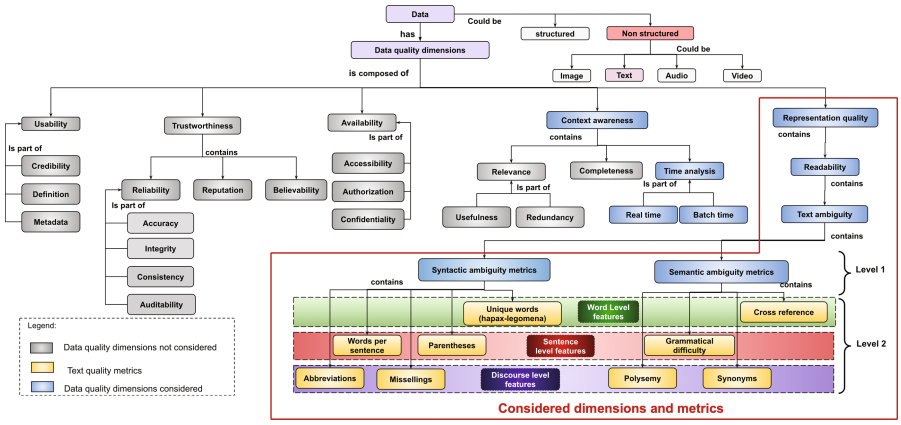

Fig. 1. Text quality dimensions

Table 1. Description of ambiguity text metrics

Level	Type	Metric	Formal definition	Formal description	Interpretation
Word	Synt	Unique words (hapax-legoma)	$\frac{1}{card(S)}\ card\,(\bigcup_{\substack{i=1 \\ w_i \neq w_j}}^{n} w_i)$	The percentage of unique words in the text	The more the number of words that have only one occurrence in a given corpus is high the more the text becomes ambiguous
	Sem	Cross reference	$\frac{1}{card(S)}\ card\,(\bigcup_{\substack{i=1 \\ w_i \in P}}^{n} w_i)$ $n \in \mathbb{N}^*$	The percentage of words that references other information[a]	This discourse-level feature increases the text ambiguity because some words could reference the same object and algorithms will not detect this reference
Sentence	Synt	Words per sentence	$\frac{1}{card(D)}\ card\,(S)$	The percentage of words per sentence	The more words the sentence contains, the more ambiguous it becomes
		Parentheses	$\frac{1}{card(S)}\ card\,(\bigcup_{\substack{i=1 \\ w_i \in D}}^{n} w_i)$ $n \in \mathbb{N}^*$	The percentage of parentheses per text	The more the sentence contains parentheses, the more it becomes ambiguous
	Sem	Grammatical difficulty	$\frac{1}{card(D)}\ card\,(\bigcup_{i=1}^{n}\bigcup_{j=1}^{k} p_{ij})$ $n \in \mathbb{N}^*$	The percentage of words that might have different positions in discourse. For example, the word "work" can be a noun or a verb in the sentence	The more words might have different positions in discourse, the more it becomes ambiguous
Discourse	Synt	Abbreviations	$abbreviation\,(W) = True \iff \forall x \in w, x \in C$	The percentage of abbreviations in text[b]	The more the discourse contains abbreviations, the more it becomes ambiguous
		Misspellings	$Misspellings\,(w) = False \iff w \in D$	The percentage of words spelled incorrectly	This indicator increases the ambiguity in texts and especially in models training
	Sem	Polysemy	$\frac{1}{card(D)} card\,(\bigcup_{i=1}^{n}\bigcup_{j=1}^{k} m_{ij})$ $n \in \mathbb{N}^*$	The percentage of words that have multiple related meanings	It presents the capacity of a word to have multiple related meanings and make it harder for prediction systems to realize that it is the same word
		Synonyms	$w\ has\ a\ synonym\ in\ S \iff \exists\, x\ \in S,\ \exists\, y \in E\,,\quad x = y$	The percentage of different words that can be synonyms in the text	This metric calculates the number of synonyms of every word in the discourse

[a]Cross reference is a notation to pertinent information at another place

[b]An abbreviation is a shortened form of a written word or phrase used to refer to names, places, companies, etc.

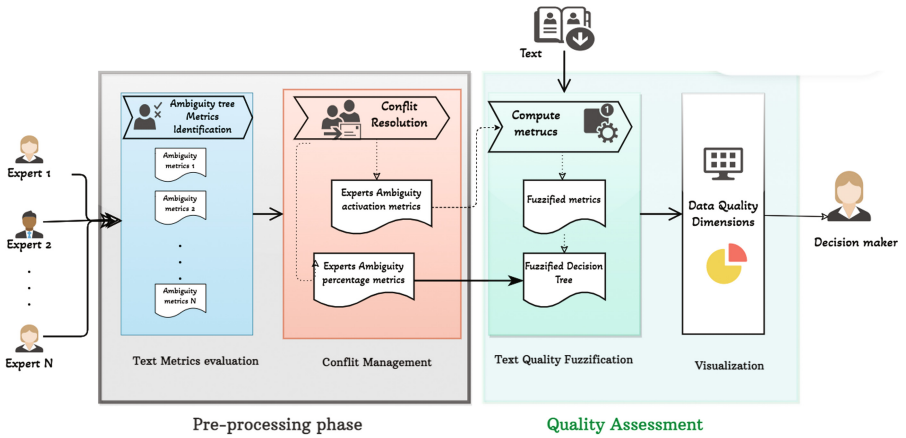

Fig. 2. Text quality evaluation approach

weighting the importance of text quality indicators. Our goal is to evaluate the importance of every indicator for the inference of a given text quality evaluation. This phase consists of two sub-phases. The first one, **"Text metrics evaluation"** is based on the knowledge of the domain experts; it deals with:

- First, the establishment of metrics' importance weighting and their relationship for high, intermediate and rudimentary levels. As the rudimentary metrics may not have the same importance for an intermediate metric for a given context, a weighting coefficient is used to reflect the relevance score of a given metric *Mh,i* to the intermediate metric *Mh+1,i*.
- Second, an activation function is defined to decide whether a metric should be activated or not. This function aims to transform the weighted metric into an output value to be fed to the next layer.

The second sub-phase is the **"Conflict management"**. Our approach is based on aggregating the weights accorded by several experts. Thus, in order to handle imprecise and conflicting experts' opinions, we apply the Evidence theory (also known as Dempster-Shafer Theory). It is a general framework for reasoning with uncertainty, with understood connections to other frameworks such as probability, possibility and imprecise probability theories. [14].

Given the problem of evaluating the text ambiguity associated with a given context, the universe of discourse Θ of the evidence theory would be seen as the set of all possible metrics for syntactic ambiguity evaluating (respectively semantic ambiguity).

The power set of Θ_{syn} noted as $2^{\Theta_{syn}}$, consists of all the subsets of Θ_{syn} such that: $\Theta_{syn} = \{\Theta_1^{syn}, \Theta_2^{syn}, \Theta_3^{syn}, \Theta_4^{syn}, \Theta_5^{syn}\}$.

Accorded weight and function activation for each metric per each expert *Ei* is expressed using evidence mass function $m_i^{syn}(x)$ known also as basic probability assignment such that:

$$m_i^{syn}(x) : 2^{\Theta_{syn}} \rightarrow [0,1] \times [0,1]$$

To access the percentage coefficient of the metric θ_i, we define the function $per(m_i)$ where:

$$\begin{array}{rcl} per(m_i) : [0,1] \times [0,1] & \rightarrow & [0,1] \\ (x,y) & \mapsto & x \end{array}$$

Moreover, to access the percentage coefficient of the metric θ_i, we define the function $act(m_i)$ where:

$$\begin{array}{rcl} act(m_i) : [0,1] \times [0,1] & \rightarrow & [0,1] \\ (x,y) & \mapsto & y \end{array}$$

$$\begin{cases} m_i^{syn}(\varnothing) = (0,0) \\ \sum\limits_{A \in 2^{\Theta_{syn}}} per(m_i^{syn}(A)) = 1 \end{cases}$$

Then, each expert is objectively weighted according to the similarity of his/her opinions with others experts opinions by means of evidence distance as given in

$$m_{1,\ldots,s}^{Aver}(X) = \frac{1}{s} \sum_{i=1}^{s} m_i(X) \tag{1}$$

where $m_i(X)$ are the representation of mass functions.

The measure of conflict between an expert Ei and all the other set of experts is:

$$conf(j,\varepsilon) = \frac{1}{n-1} \sum_{e=1}^{n} conf(j,e) \tag{2}$$

Finally, adjusted scores are combined to generate the weighting coefficient using the Dempster's combination rule for combining two or more belief functions [15].

3.2.2 Quality Assessment Based on Fuzzy Decision Tree Inference

To assess TQ ambiguity, we investigate the hierarchical representation of metrics and fuzzy logic inference. We need to extend different fuzzified values of rudimentary metrics (a subset U) to intermediate or high-level metrics (which are fuzzy subset). Thus, we chose the extension principle that is in fact a special case of the compositional rule of inference.

The extension principle, described by [14] is a general method for extending crisp mathematical concepts to address fuzzy quantities. It is particularly useful in connection with the computation of linguistic variables, the computing of linguistic probabilities, arithmetic of fuzzy numbers, etc. We applied this theory to deduce metrics value in the higher level of ambiguity tree. Thus, the extension principle is defined:

$$\mu_B(y) = sup\{min(\mu_\phi(x,y), \mu_A(x)/x \in E\} \tag{3}$$

where:

- A is the set which includes syntactic ambiguity metrics M_1, M_2, .., M_n, C is the set which includes semantic ambiguity metrics M_1, M_2, .., M_n for a given level.
- B is the set which includes fuzzy data type used to represent the text ambiguity degree of a given text A="Very High ambiguity", "High ambiguity", "Normal ambiguity", "Low ambiguity", "Very Low ambiguity".
- ϕ is a function that associates x $\in$ A to y $\in B$, $\phi(x) = y$

To explain the fuzzification part, a metric M_i and a threshold value $M_{th,i}$ is fixed by experts for a text T in a given context. The max between $(M_i - M_{th,i})$ and 0 is considered. Then, the determined value is treated by a sigmoid function to compute the ambiguity level. For example, if $M_1 = 0.3$ and $M_{th_1} = 0.1$. The result is: $max(0.3 - 0.1, 0) = 0.2$. Finally, passing by the sigmoid function, the obtained result is $\mu_{M1}(T) =$ Very Low.

4 Experimental Study

This section presents the data collection and acquisition process and quality computing result before evaluating the quality model.

4.1 Data Collection and Acquisition

We leverage a meticulously curated dataset sourced from Kaggle [16] that is structured with two pivotal columns: "text", which contains the text of the tweets, and "sentiments", which indicates the sentiment of the tweets and ranges between -1, 0, and 1. To enrich our data repository, we seamlessly integrate the Tweepy Python library to our developed interface allowing experts to customize the weight of each metric and to choose the subject of scrapped data as shown in the Fig. 3.

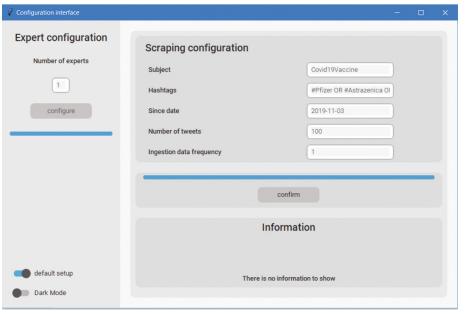

Fig. 3. Configuration interface

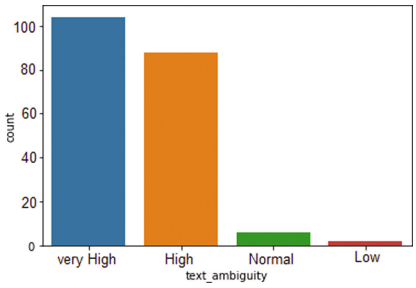

Fig. 4. Evaluation of text ambiguity

Table 2. Evaluation of forecasting models

Model	RMSE
Polynomial regression	0.0046
Holt's Linear	0.0032
AUTO ARIMA	0.00054

Table 3. Evaluation of sentiment analysis models

Model	Accuracy	f1-score
LSTM	84.65 %	0.845
XGBoost	84.03 %	0.836
Random Forest	80.09 %	0.797
Naive Bayes	70.46 %	0.685

Table 4. The effect of text quality on sentiment analysis models

Metrics	Before[a]		After[b]	
	Accuracy	F1-Score	Accuracy	F1-Score
Random Forest	40%	0.333	43.34%	0.351
Naive Bayes	48.5%	0.452	53%	0.472
XGBoost	32%	0.357	34.66 %	0.379

[a]Before eliminating very High ambiguous Data.
[b]After eliminating very High ambiguous Data.

4.2 Quality Computing Results

The goal of the experiment is to illustrate how the quality of a Twitter stream can be assessed using the dimension and metrics presented in Sect. 3. Sentiment analysis in Covid-19 vaccine is token as a case study to illustrate the usefulness of our approach.

Sentiment Analysis Models Evaluation. The Table 3 presents the evaluation results of 4 sentiment analysis models which show that LSTM has the best accuracy and f1-score.

Forecasting Models Evaluation. Three forecasting models with this data were trained and the evaluation results are shown in Table 2: AUTO ARIMA forecasting model has the lowest value of RMSE (Root Mean Square Error).

4.3 Quality Model Evaluation

We evaluated the quality of 200 texts which present more than 50% of very high ambiguity as shown in Fig. 4. We trained 3 ML algorithms with and without very high ambiguous data. The obtained results, shown in Table 4, ensure that TQ is one of necessary exigences to get better results. Despite the limited quantity of texts used for training sentiment analysis models (which accounts for the relatively low accuracy and F1-score values), the removal of high ambiguous data induce an improvement in the performance of the sentiment analysis models.

5 Conclusion and Future Directions

In light of the growing concern surrounding data quality in sentiment analysis for decision-making, this research presents an automatic text quality approach that can be scalable and applicable to machine learning applications within different contexts. By leveraging the principles of the data quality model, evidence theory, and fuzzy logic reasoning, we can improve the accuracy and reliability of sentiment analysis algorithms. The key contributions of this research are as follows: (1) a hierarchical decomposition of the text quality model tree to address both syntactic and semantic ambiguity, (2) contextual weighting of metrics by experts and conflict management, and (3) fuzzified quality inference by integrating weighted metrics evaluated at both low-level and high-level measurements. We believe that this proposal can be gradually enhanced by integrating additional DQ dimensions and metrics. Furthermore, the system architecture has the potential to be enriched with intelligent features and components that facilitate the derivation of contextual recommendations.

References

1. Arolfo, F., Rodriguez, K.C., Vaisman, A.: Analyzing the quality of twitter data streams. Inf. Syst. Front. 1–21 (2020)
2. Wand, Y., Wang, R.Y.: Anchoring data quality dimensions in ontological foundations. Commun. ACM **39**(11), 86–95 (1996)
3. Handbook, A.: From contract drafting to software specification: linguistic sources of ambiguity (2003)
4. Khezri, R.: Automated detection of syntactic ambiguity using shallow parsing and web data (2017)
5. Ali, K., Dong, H., Bouguettaya, A., Erradi, A., Hadjidj, R.: Sentiment analysis as a service: a social media based sentiment analysis framework. In: 2017 IEEE International Conference on Web Services (ICWS), pp. 660–667. IEEE (2017)
6. Pollacci, L., SSîrbu, A., Giannotti, F., Pedreschi, D., Lucchese, C., Muntean, C.I.: Sentiment spreading: an epidemic model for lexicon-based sentiment analysis on twitter. In: Esposito, F., Basili, R., Ferilli, S., Lisi, F. (eds.) AI*IA 2017. LNCS, vol. 10640, pp. 114–127. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70169-1_9
7. Alamoodi, A.H., et al.: Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: a systematic review. Expert Syst. Appl. **167**, 114155 (2021)
8. Soto, A., et al.: Data quality challenges in twitter content analysis for informing policy making in health care (2018)
9. Murshed, B.A.H., Abawajy, J., Mallappa, S., Saif, M.A.N., Al-Ghuribi, S.M., Ghanem, F.A.: Enhancing big social media data quality for use in short-text topic modeling. IEEE Access **10**, 105328–105351 (2022)
10. Suanmali, L., Salim, N., Binwahlan, M.S.: Fuzzy logic based method for improving text summarization. arXiv preprint arXiv:0906.4690 (2009)
11. Arruda, N., et al.: A fuzzy approach for data quality assessment of linked datasets. In: International Conference on Enterprise Information Systems, vol. 1, pp. 399–406. SciTePress (2019)

12. Cichy, C., Rass, S.: Fuzzy expert systems for automated data quality assessment and improvement processes. In: EKAW (Posters & Demos), pp. 7–11 (2020)
13. Salvatore, C., Biffignandi, S., Bianchi, A.: Social Media and Twitter Data Quality for New Social Indicators. Soc. Indicat. Res. **156**(2), 601–630 (2021). ISSN 1573-0921
14. Zadeh, L.A., Klir, G.J., Yuan, B.: Fuzzy sets, fuzzy logic, and fuzzy systems. Adv. Fuzzy Syst. Appl. Theory **6** (1996)
15. Shafer, G.: Dempster's rule of combination. Int. J. Approximate Reasoning **79**, 26–40 (2016)
16. Nasreen Taj, M.B., Girisha, G.S.: Insights of strength and weakness of evolving methodologies of sentiment analysis. Glob. Transit. Proc. **2**(2), 157–162 (2021)