SA-Q: Observing, Evaluating, and Enhancing the Quality of the Results of Sentiment Analysis Tools

Wissam Maamar-Kouadri
Université Paris Cité, LIPADE
wissam.maamar-kouadri@etu-paris.fr

Salima Benbernou
Université Paris Cité, LIPADE
salima.benbernou@u-paris.fr

Mourad Ouziri
Université Paris Cité, LIPADE
mourad.ouziri@u-paris.fr

Themis Palpanas
Université Paris Cité, LIPADE &
French University Institute (IUF)
themis@mi.parisdescartes.fr

Iheb Ben Amor
IMBA Consulting
iheb.benamor@imba-consulting.com

ABSTRACT

Sentiment analysis has received constant research attention due to its usefulness and importance in different applications. However, despite the research advances in this field, most current tools suffer in prediction quality due to the inconsistencies in their results, i.e., intra- and inter-tool inconsistencies. This demonstration proposes a system for the evaluation of sentiment analysis quality namely SA-Q. The system allows the evaluation of inconsistency in sentiment analysis tools, the resolution of the inconsistency using state-of-the-art methods and the recommendation of relevant sentiment analysis tool for any type of data set provided by the attendees. It allows the attendees to compare the tools. Moreover, we demonstrate that SA-Q evaluates the consistency of tools on two levels (intra-tool and inter-tool). Through various scenarios, we showcase the challenges of inconsistency resolution, demonstrate the usefulness of the proposed system and the recommendations that can be given to the attendees for their datasets. We demonstrate that SA-Q system has practical utility in many areas of industrial applications for better decision making. This demonstration shows promising research areas for data management, NLP, and machine learning communities by adopting and drawing inspiration from truth inference methods to create more robust tools and improve the tool's scalability.

PVLDB Reference Format:
Wissam Maamar-Kouadri, Salima Benbernou, Mourad Ouziri, Themis Palpanas, and Iheb Ben Amor. SA-Q: Observing, Evaluating, and Enhancing the Quality of the Results of Sentiment Analysis Tools. PVLDB, 15(12): 3658-3661, 2022.
doi:10.14778/3554821.3554868

1 INTRODUCTION

With the growing popularity of social media, people are increasingly sharing their opinion online about products, services, and entities, making the sentiment analysis of social media content crucial in organizations' decision-making process. Therefore, several studies have been interested in sentiment analysis [9, 12, 14, 15], which can be defined as the process of automatically extracting the polarity, i.e., positive, negative, or neutral, from the document (text). Nevertheless, despite the advances made in this research area, sentiment analysis is still a challenging task due to the complexity and variety of natural language where the same idea can be expressed and interpreted using different text. Let us illustrate this issue by considering the two texts (documents): *(d1) Donald Trump softens the tone on Chinese investments*; *(d2) Trump drops new restrictions on China investment*. We notice that although the two documents are structured differently, they are, in fact, semantically equivalent paraphrases because they convey the same meaning. Many research works [5, 8] have agreed that semantically equivalent documents should have the same polarity. However, through the intensive experiments conducted in our work [13] and in [10], we notice that most sentiment analysis tools assign different polarities to semantically equivalent documents. Hence, considering *intra-tool inconsistency*, where the sentiment analysis tool attributes different polarities to the semantically equivalent documents, and *inter-tool inconsistency*, where different sentiment analysis tools attribute different polarities to the same document. For instance, the analysis of documents d1 and d2 using the algorithm [2] returns positive and negative polarities, respectively, leading to an intra-tool inconsistency. On the other hand, analyzing document d2 using the tools [15] and [2] yields neutral and positive polarities, respectively, causing an inter-tool inconsistency. Inconsistencies signify that the tool makes prediction errors and that at least one tool has given an incorrect polarity which makes them harmful and leads to poor business decisions. In this demonstration, we present the SA-Q (Sentiment Analysis Quality) system that uses the findings of our work [13] to evaluate tools' quality and implements several state-of-the-art methods for inconsistency resolution (i.e. PM [1], and ZC [4]). The system is a web application that allows attendees to perform sentiment analysis, evaluate sentiment analysis tools' consistency, resolve the inconsistency, visualize the results through a dashboard and recommend appropriate tool(s) for a given data type. We showcase the usefulness of SA-Q and the effect of resolving tool inconsistencies using real datasets and the benchmark we developed through various scenarios. The demonstration shows the usefulness of the system to help companies to select a tool or a combination of tools to perform sentiment analysis.

Proceedings of the VLDB Endowment, Vol. 15, No. 12 ISSN 2150-8097.
doi:10.14778/3554821.3554868

2 PROPOSED APPROACH

In this section, we recall definitions of sentiment consistency from [13], then, we present the set of sentiment analysis tools to be used for the evaluation, the set of methods to detect the inconsistency in the sentiment analysis tools, and the evaluation measures.

2.1 Definitions

2.1.1 *Analogical Set.*

It is a set of semantically equivalent documents: $A_l = \{d_1, \ldots, d_n\}$ s.t: $\forall d_i, d_j \in A_l \ \ , d_i \overset{s}{\Longleftrightarrow} \ \ d_j$ where $\overset{s}{\Longleftrightarrow}$ denotes semantic equivalence, and d_i and d_j are documents.

2.1.2 *Sentiment Consistency.*

For each dataset D and polarity functions set Γ, we define sentiment consistency as the two rules 1) and 2):

1) Intra-tool consistency. It assesses the contradiction of tools when considered individually. In other words, there is an Intra-tool inconsistency in a tool P_{t_k} if it assigns different polarities to two analogical documents d_i and d_j:

$$\forall A_l \ \forall d_i, d_j \in A_l \ \forall P_{t_k} \in \Gamma, \ P_{t_k}(d_i) = P_{t_k}(d_j) \tag{1}$$

2) Inter-tool consistency. It assesses the contradiction between the tools. In other words, there is an Inter-tool inconsistency between the tools P_{t_k} and $P_{t'_k}$ if they assigns different polarities to a given document d_i:

$$\forall A_l \ \forall d_i \in D \ \forall P_{t_k}, P_{t'_k} \in \Gamma \quad P_{t_k}(d_i) = P_{t'_k}(d_i) \tag{2}$$

2.2 Sentiment Analysis Tools and Benchmark

In this demonstration, we use a set of ten representative methods from the main categories of sentiment analysis methods: lexicon-based methods, lexicon-based methods with rules, and machine learning methods. We use **SentiWordnet** [7], for sentiment analysis as a representative method for the lexicon-based category, **Vader [9]** and **SenticNet** [3], which are lexicon with a set of rules as representative methods for rule-based approaches. In the learning-based methods, we use three machine learning tools: **RecNN** [15] that learns word embeddings, **Text_ CNN** [12] that uses a pre-trained word embedding methods including BERT, Word2vec and Glove, and **Char_CNN** [6], which uses two levels of embedding character and word embedding. These methods are widely used in both industry and literature and are robust. We also use very recent tools **SentiBERT** [16] a variant of BERT that captures better compositional sentiment semantics and **SentiLARE** [11] a novel pre-trained model that proposes a word-level linguistic knowledge from SentiWordNet via context-aware sentiment attention. We refer to sentiment analysis tools by their polarity functions, i.e., SenticNet as $P_{senticnet}$. In addition to the datasets to be provided by the attendees, we use our own benchmark comprising five datasets, augmented with paraphrases to evaluate the intra-tool inconsistency and whose quality was refined using the algorithm in [13]. Our benchmark covers the most used domains by companies such as costumer reviews, movie reviews, news and tweets.

2.3 Inconsistency Resolution Methods

To resolve the inconsistencies in sentiment analysis tools, we have suggested in our paper [13] to use state-of-the-art methods for truth inferences. Furthermore, since truth inference methods are classified to probabilistic model, direct computing, and optimization [17], we use the following methods:

Majority voting (MV) represents the direct and trivial computation to resolve inconsistency and infers truth.

$p^*(d_i) = \underset{\Omega \in \{+,0,-\}}{argmax} \ \sum_{p_{t_k} \in F} \mathbb{1}_{\{p_{t_k}(d_i) = \Omega\}}$

Zencrowd (ZC) is the primary method for inconsistency resolution using the probabilistic graphical model.

$Pr(p_{t_k}(d_i)|q_{t_k}, p^*(d_i)) = q_{t_k}^{\mathbb{1}_{p_{t_k}(d_i) = p^*(d_i)}} .(1 - q_{t_k})^{\mathbb{1}_{p_{t_k}(d_i) \neq p^*(d_i)}}$

PM represents the primary method that uses optimization to learn the quality of tools and infer the truth by optimizing the following objective

$\min q^{t_k}, p^*(d_i) \sum_{t_k \in F} q^{t_k} \sum d(p^*(d_i), p_{t_k}(d_i))$

with $p^*(d_i)$ is the inferred polarity of the document (truth), q^{t_k} the quality of the tool t_k, $\Omega \in \{+, 0, -\}$ represents the polarity, and $d(.,.)$ is the distance between two polarities.

2.4 Metrics

- We evaluate *Accuracy* by calculating the rate of correctly predicted polarities (P^*) compared to the golden polarity (P_h) as follows: $Accuracy = \frac{\sum_{i=0}^{n} \mathbb{1}_{(P_h(d_i) = P^*(d_i))}}{n}$
- We evaluate the *intra-tool* inconsistency rate for each document d_i in the analogical set A and sentiment analysis tool t_k as the proportion of documents d_j that have a different polarity than $P_{t_k}(d_i)$ with regards to the tool t_k. We write

$$\begin{gathered}\forall d_i \in A, P_{t_k} \in \Gamma, inc_{in}(d_i, P_{t_k}) = \frac{card(S)}{n-1} \\ \text{s.t } S = \{d_j \in A | P_{t_k}(d_i) \neq P_{t_k}(d_j)\}\end{gathered} \tag{3}$$

The *intra-tool* inconsistency rate of an analogical set A is the mean of different intra-tool inconsistency rates of its documents:

$$inc_{in}(A, P_{t_k}) = \frac{\sum_{j=1}^{n} inc_{in}(d_j, P_{t_k})}{n}, n = card(A) \tag{4}$$

- We measure the *inter-tool* inconsistency rate for each tool t_k and document d_j as the rate of tools $t_{k'}$ that give different polarities to the document d_j than P_{t_k}. We write:

$$\begin{gathered}\forall P_{t_k} \in \Gamma, \forall d_j \in A, inc_{inter}(d_j, P_{t_k}) = \frac{card(S')}{m-1} \\ \text{s.t } S' = \{P_{t'_k} \in \Gamma | P_{t'_k}(d_j) \neq P_{t_k}(d_j)\}\end{gathered} \tag{5}$$

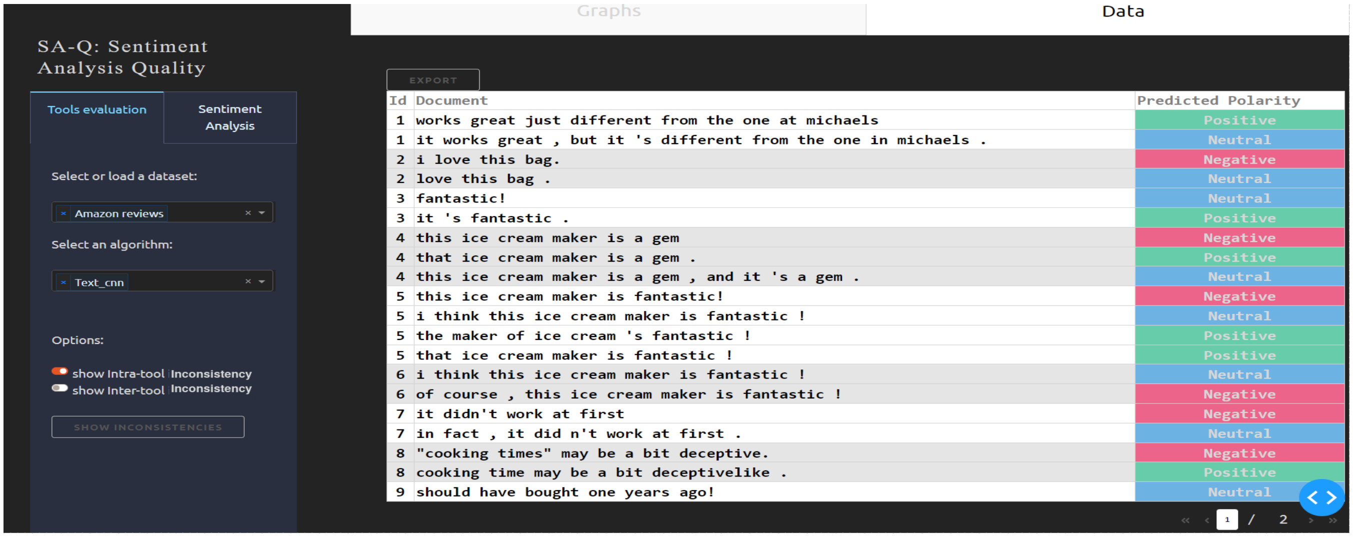
The *inter-tool* inconsistency rate in the set Γ is the mean of inconsistency rates of the different tools:

$$inc_{inter}(d_j, \Gamma) = \frac{\sum_{k=1}^{m} inc_{inter}(d_j, P_{t_k})}{m}, m = card(\Gamma) \tag{6}$$

Γ is the set of all polarity functions, P_{t_k} and $P_{t'_k}$ polarity functions, A is an analogical set, d_i and d_j are documents.

3 SA-Q OVERVIEW

SA-Q is a web application enables users to perform sentiment analysis using ten state of the art tools, evaluating their quality, i.e., accuracy, consistency, scalability, resolving the inconsistency between them for an integrated prediction, and recommending a sentiment analysis method for a given dataset. SA-Q's architecture



SA-Q: Sentiment Analysis Quality

Tools evaluation | Sentiment Analysis

Select or load a dataset: Amazon reviews

Select an algorithm: Text_cnn

Options:

● show Intra-tool Inconsistency

○ show Inter-tool Inconsistency

SHOW INCONSISTENCIES

Graphs | Data

EXPORT

Id	Document	Predicted Polarity
1	works great just different from the one at michaels	Positive
1	it works great , but it 's different from the one in michaels .	Neutral
2	i love this bag.	Negative
2	love this bag .	Neutral
3	fantastic!	Neutral
3	it 's fantastic .	Positive
4	this ice cream maker is a gem	Negative
4	that ice cream maker is a gem .	Positive
4	this ice cream maker is a gem , and it 's a gem .	Neutral
5	this ice cream maker is fantastic!	Negative
5	i think this ice cream maker is fantastic !	Neutral
5	the maker of ice cream 's fantastic !	Positive
5	that ice cream maker is fantastic !	Positive
6	i think this ice cream maker is fantastic !	Neutral
6	of course , this ice cream maker is fantastic !	Negative
7	it didn't work at first	Negative
7	in fact , it did n't work at first .	Neutral
8	"cooking times" may be a bit deceptive.	Negative
8	cooking time may be a bit deceptivelike .	Positive
9	should have bought one years ago!	Neutral

« ‹ 1 / 2 › »

Figure 1: Display inconsistencies in SA-Q

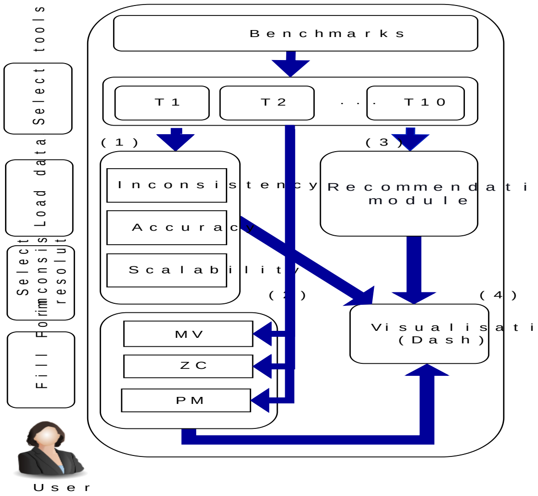


Figure 2: SA-Q system architecture (1) evaluation module (2) inconsistency resolution module (3) recommendation module (4) visualisation module. T1=SentiWordnet, T2= Vader, T3= SenticNet, T4= RecNN, T5= Text_CNN + word2vec, T6= Text_CNN + Glove, T7= Text_ CNN + Bert, T8= Char_CNN, T9= SentiBERT, and T10=SentiLARE

is presented in Figure 2. The GUI of SA-Q allows users to load datasets, select tools to evaluate, select the inconsistency resolution methods, and fill the recommendation form.

[GUI] interacts directly with the modules: inconsistency evaluation (1), inconsistency resolution (2), and recommendation (3). When the user wants to evaluate tool inconsistencies, the GUI sends the loaded data to tools for polarity extraction, then transfers the resulting logs to the inconsistency evaluation module. After that, it transfers the evaluation results to the visualization module. When the user chooses to resolve the inconsistencies, SA-Q sends the loaded dataset to the selected tools for polarity extraction and then transfers the logs to the selected inconsistency resolution method to learn the tools' confidence and resolve the inconsistency. Then it sends the logs to the visualization module, where it aggregates and displays the results.

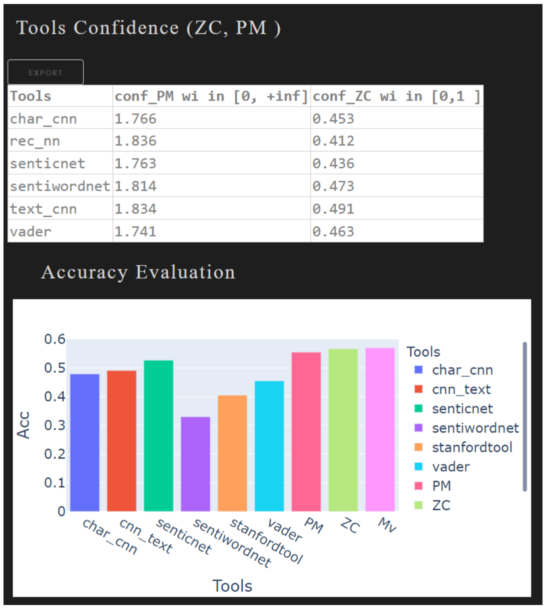
[Visualisation Module] SA-Q offers a visualization module that facilitates the comparison of tools quality and the performance of inconsistency resolution methods. It offers two types of visualization: aggregated visualization with graphs and text visualization that allows showing inconsistencies.

4 DEMONSTRATION SCENARIOS

The demonstration has the following goals: (1) observe that inconsistencies are frequent on the most known sentiment analysis tools (2) show the impact of resolving different inconsistencies on tools accuracy (3) show the enhancement obtained when resolving inconsistency using state-of-the-art methods (4) involve the participants in different scenarios (5) demonstrate the usefulness of our guideline in choosing relevant tools.

[Scenario 1: Consistency Evaluation] This scenario aims to show how inconsistent tools are and the efficiency of SA-Q in the evaluation, and the identification of inconsistencies.

The participant will evaluate the consistency of tools given a set of semantically equivalent documents sampled from the benchmark we developed in [13]. This benchmark contains 92995 documents that initially come from five publicly available datasets (news headlines, amazon reviews, movie reviews, US airlines tweets, and first GOP debate tweets), augmented with paraphrases cleaned using a heuristic. The participant can observe how changing punctuation or replacing a word with its synonym will lead the tool to miss-classify the documents' polarity. An example of inconsistencies is presented in Figure 1. After that, we run the ten sentiment analysis tools implemented in the system on the benchmark and ask the participants to identify the inconsistent cases on a large dataset manually. Then, we show how SA-Q can help identify and



Tools Confidence (ZC, PM)

Tools	conf_PM wi in [0, +inf]	conf_ZC wi in [0,1]
char_cnn	1.766	0.453
rec_nn	1.836	0.412
senticnet	1.763	0.436
sentiwordnet	1.814	0.473
text_cnn	1.834	0.491
vader	1.741	0.463

Accuracy Evaluation

Figure 3: Accuracy optimization after resolving inconsistency and tools confidence.

aggregate the inconsistencies by giving a global vision of tools inconsistency. The user will also evaluate the scalability of tools in time and dataset size. Finally, we evaluate the accuracy of the tools and demonstrate that even accurate tools may be inconsistent.

[Scenario 2: Inconsistency Resolution] In this scenario, we show how truth inference methods can resolve the inconsistencies produced by sentiment analysis tools and the accuracy improvement that can be obtained. First, we run a set of selected sentiment analysis tools implemented in SA-Q on an example and show their accuracy; then, we ask the participants to resolve the inconsistencies manually and point out the obtained accuracy improvement. In the second step of this scenario, we load a large dataset containing 1580 documents and challenge the attendees to identify and resolve the inconsistencies manually and show the difficulty of this task. After that, we run all inconsistency resolution methods included in SA-Q, .i.e, ZC, PM, and MV, and compare their performance for inconsistency resolution and tools ranking. An example of accuracy improvement obtained after resolving inconsistencies using six methods among the ten included in SA-Q is displayed in Figure 3.

[Scenario 3: Usefulness of Truth Inference Methods] This scenario demonstrates the usefulness of considering different tools and ranking them by their quality when resolving inconsistencies. We first demonstrate the accuracy improvement obtained by resolving inconsistencies using majority voting. We explain majority voting limits using an example and show how tools ranking methods (ZC and PM) outperform majority voting on the dataset. We show attendees the difference between the inconsistency resolution methods and specify the usefulness of each technique through examples.

[Scenario 4: Tools' recommendation] In this scenario, we demonstrate how we can use our guideline, described in [13], to choose a sentiment analysis tool. First, we ask the attendees to load a dataset from the benchmark and fill a form about the data characteristics such as the document's length, the data source (social media data, news headlines, reviews), data type (financial, medical, political data, or product reviews). Then, based on the filled form, the dataset, and the inconsistency score, SA-Q recommends the appropriate sentiment analysis tool(s).

ACKNOWLEDGMENTS

We are grateful to H.Lahmar for his help on this paper.This work is partially supported by IMBA Consulting and ANRT French Research Fund for the project e-reputation Number 2018/0576.

REFERENCES

[1] B. I. Aydin, Y. S. Yilmaz, Y. Li, Q. Li, J. Gao, and M. Demirbas. Crowdsourcing for multiple-choice question answering. In *Twenty-Sixth IAAI Conference*, 2014.
[2] E. Cambria, Y. Li, F. Z. Xing, S. Poria, and K. Kwok. Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis. In *Proc of the 29th ACM CIKM*, pages 105–114, 2020.
[3] E. Cambria, S. Poria, D. Hazarika, and K. Kwok. Senticnet 5: discovering conceptual primitives for sentiment analysis by means of context embeddings. In *Proceedings of AAAI*, 2018.
[4] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st international conference on World Wide Web*, pages 469–478, 2012.
[5] H. Ding and E. Riloff. Weakly supervised induction of affective events by optimizing semantic consistency. In *Proc of the 32th AAAI*, pages 5763–5770, 2018.
[6] C. Dos Santos and M. Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78, 2014.
[7] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, volume 6, pages 417–422. Citeseer, 2006.
[8] G. Fu, Y. He, J. Song, and C. Wang. Improving chinese sentence polarity classification via opinion paraphrasing. In *Proc of The 3rd CIPS-SIGHAN*, pages 35–42, 2014.
[9] C. H. E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proc of the 8th ICWSM-14.*, 2014.
[10] M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, 2018.
[11] P. Ke, H. Ji, S. Liu, X. Zhu, and M. Huang. Sentilare: Sentiment-aware language representation learning with linguistic knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6975–6988.
[12] Y. Kim. Convolutional neural networks for sentence classification. *arXiv:1408.5882*, 2014.
[13] W. M. Kouadri, M. Ouziri, S. Benbernou, K. Echihabi, T. Palpanas, and I. B. Amor. Quality of sentiment analysis tools: The reasons of inconsistency. *Proc. VLDB Endow.*, 14(4):668–681, 2020.
[14] A. Severyn and A. Moschitti. Twitter sentiment analysis with deep convolutional neural networks. In *Proc of the 38th SIGIR*, pages 959–962. ACM, 2015.
[15] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc of EMNLP*, pages 1631–1642, 2013.
[16] D. Yin, T. Meng, and K. Chang. Sentibert: A transferable transformer-based architecture for compositional sentiment semantics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3695–3706.
[17] Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng. Truth inference in crowdsourcing: Is the problem solved? *Proc of the VLDB Endow*, 10(5):541–552, 2017.