Engineers, Aware! Commercial Tools Disagree on Social Media Sentiment: Analyzing the Sentiment Bias of Four Major Tools

Engineers, Aware! Commercial Tools Disagree on Social Media Sentiment

Analyzing the Sentiment Bias of Four Major Tools

SOON-GYO JUNG, Qatar Computing Research Institute, Hamad bin Khalifa University, Qatar
JONI SALMINEN, University of Vaasa, Finland
BERNARD J. JANSEN, Qatar Computing Research Institute, Hamad bin Khalifa University, Qatar

Large commercial sentiment analysis tools are often deployed in software engineering due to their ease of use. However, it is not known how accurate these tools are, and whether the sentiment ratings given by one tool agree with those given by another tool. We use two datasets – (1) NEWS consisting of 5,880 news stories and 60K comments from four social media platforms: Twitter, Instagram, YouTube, and Facebook; and (2) IMDB consisting of 7,500 positive and 7,500 negative movie reviews – to investigate the agreement and bias of four widely used sentiment analysis (SA) tools: Microsoft Azure (MS), IBM Watson, Google Cloud, and Amazon Web Services (AWS). We find that the four tools assign the same sentiment on less than half (48.1%) of the analyzed content. We also find that AWS exhibits neutrality bias in both datasets, Google exhibits bi-polarity bias in the NEWS dataset but neutrality bias in the IMDB dataset, and IBM and MS exhibit no clear bias in the NEWS dataset but have bi-polarity bias in the IMDB dataset. Overall, IBM has the highest accuracy relative to the known ground truth in the IMDB dataset. Findings indicate that psycholinguistic features – especially affect, tone, and use of adjectives – explain why the tools disagree. Engineers are urged caution when implementing SA tools for applications, as the tool selection affects the obtained sentiment labels.

CCS Concepts: • **Computing methodologies** → **Machine learning approaches**; • **General and reference** → **Evaluation**.

Additional Key Words and Phrases: sentiment analysis, evaluation, bias, agreement

1 INTRODUCTION

Sentiment analysis (SA) is the use of computational techniques for understanding attitudes, affects, and opinions expressed by users in text materials, such as user-generated content on the Internet [16, 57]. This goal of understanding online sentiment of users is crucial for various stakeholder groups responsible for user-centric decisions, such as the engineering community that uses black-box SA services for integrating social media analysis functions into their applications.

In the last decade, there has been an explosion of research on different types of SA [36, 56] fueled by volumes of sentiment rich text accessible on social media. This content includes blogs, review sites, forums, tweets, and other user generated content (UGC) critical to decision-making for companies, researchers, policymakers, and other stakeholders interested in social media analysis. For example, news and media organizations are interested in measuring the reactions of their audiences in order to report on stories in a non-polarizing way [3], scholars in Web and computational social sciences seek to understand the online behaviors of users [36, 40] to gauge reaction to systems and interfaces, whereas policy makers desire to poll the public sentiment about various topics [31] to base their decisions on *vox populis*, as well as to monitor reactions to political decisions once in effect [10]. Finally, commercial organizations, including startups and Fortune 500 companies, are keenly motivated to track the reception of new product launches, customer satisfaction,

Authors' addresses: Soon-gyo Jung, sjung@hbku.edu.qa, Qatar Computing Research Institute, Hamad bin Khalifa University, Doha, Qatar; Joni Salminen, jonisalm@uwasa.fi, University of Vaasa, Vaasa, Finland; Bernard J. Jansen, bjansen@hbku.edu.qa, Qatar Computing Research Institute, Hamad bin Khalifa University, Doha, Qatar.

correlation of online sentiment and stock prices, and similar phenomena via online/offline SA [12, 41]. All these use cases require robust SA tools.

Major software corporations provide such SA tools via Application Programming Interfaces (APIs) to address these vital stakeholder needs for interactive systems. These API-based services are popular, enabling software developers to quickly deploy SA functionalities in their applications. In 2019, the combined market share of these four platforms was 60% of the cloud-services market[1]. On average, there are 563M monthly visits (combined all four platforms) to the cloud providers' websites (statistics from April 2020[2]). These figures reflect the popularity of API-based services in software development. Through adopting APIs in software systems, the popular SA tools can influence social processes, such as decision making about public policy, products, politics, finances, social services, and employment [16]. If the applied SA tool is biased, this would mean downstream decisions contingent on sentiment information can become biased as well – as such, the bias in SA tools poses a concern going beyond technology. Goncalves et al. [20] mention the example of news on an airline crash being considered as "positive" by several SA tools. Jung et al. [25] show that demographic bias can originate from API-based image classification.

In order to present accurate insights about the sentiment of people, API-based SA tools need to provide reasonably impartial calculations of the sentiment of a given comment or content without systematic biases towards a specific sentiment label. Despite this being commonly known and understood in the literature [12, 16], we could locate no previous study that specifically focuses on analyzing the sentiment labels given by the four major commercial SA tool providers: Amazon Web Services (AWS), Google Cloud, IBM Watson, and Microsoft Azure (MS).

In particular, while researchers have applied different methods to understand the sentiment of online users [13, 22, 28], it is unclear how well different SA tools agree or disagree about the sentiment of social media content. Several risks exist in this regard. For example, a tool may systematically label comments towards more negative or positive polarity, risking biasing the information shown to end users of applications and systems that rely on the SA tools. Therefore, it is essential to investigate if the results provided by these SA tools are consistent and unbiased.

In this research, we analyze if four popular SA tools include a systematic bias in terms of assigning sentiment (positive/neutral/negative) to online content and comments of online news stories from a major international news organization. Our research questions (RQs) are as follows:

RQ1: How well do different SA tools agree on the exact same content?

RQ2: Are SA tools systematically biased in assigning sentiment labels for the exact same content?

RQ3: What linguistic features cause disagreement among the tools?

We define 'sentiment bias' as a systematic deviation of the sentiment value given by a SA tool compared to the average sentiment given by all the tested tools. Note that we define sentiment rating as an inherently biased or subjective task – for this reason, we are not interested in measuring bias in terms of deviation from "ground truth", but rather as a deviation from the sentiment labels (i.e., positive, negative, neutral) given by the entire set of tools. As the content labeled by all the SA tools is constant, investigating the ratings' differences inform us if and how a tool's ratings are skewed toward a particular sentiment class. This limitation is critically important for the reader to understand: *by "biased", we mean biased in relation to each other, not in relation to any ground truth*[3]. The discussion section contains more specifics on this research choice.

[1]https://wire19.com/cloud-services-comparison-tool-aws-vs-google-vs-ibm-vs-microsoft
[2]https://www.similarweb.com
[3]Note that we also inspect ground truth values for the second dataset we obtained, in order to provide recommendations for engineers on which SA tools to select.

In addition, we analyze the sentiment tendencies to understand how users tend to react to news stories on different platforms. Findings enhance understanding whether the platforms themselves are environments that affect the quality of comments. Inaccurate or biased sentiment values given by an algorithm can misrepresent public opinions about a topic or propagate social biases against certain groups based on their sociodemographic characteristics [16]. Consequently, awareness of the SA tools' limitations is needed to more appropriately analyze social media discussions.

2 RELATED WORK

2.1 Principles of Sentiment Analysis

SA has evolved into a burgeoning field uniting computer science and linguistics. It involves the contextual mining of text to extract and classify positive and negative opinions, emotions, and appraisals. The field uses natural language processing (NLP), text analysis, and statistics to analyze customer and user sentiment in texts. The principal aim of SA is to determine the polarity and strength of polarity of a subjective text [18, 46]. Typically, this involves three levels of analysis. At the document level, the semantic orientation of text is classified as negative, positive, or neutral. At the sentence-level, sentences are classified in the same manner. At the aspect-level analysis, the text is broken down into aspects (topics), and each one is assigned a sentiment level [4].

Overall, SA is a complex process involving multiple steps to prepare, classify, and interpret data. Three main approaches are typically used: (a) lexicon-based; (b) machine learning (ML); (c) and hybrid [4]. The lexicon-based approach involves calculating sentiment from the semantic orientation (polarity and strength) of words or phrases in a text [47, 51]. This method uses a sentiment lexicon or dictionary, a list of lexical features (words, phrases, etc.) that have been assigned a polarity value. This list is then used to calculate a score for the input text's polarity and/or sentiment. Hybrid approaches combine lexicon-based and ML-based techniques. Merging the two approaches can help offset some of the disadvantages of the ML- and lexicon-based methods to improve sentiment classification performance.

2.2 Bias in Sentiment Analysis

There is a small but growing body of literature on bias in SA systems. This line of research explores how biases are perpetuated through SA methods, producing output that is inaccurate, unhelpful, and even harmful to business, political or economic decision-making.

2.2.1 Origins of Bias. Bias in SA can arise throughout the process and from a variety of sources. Several authors discuss the implications of introducing biased data into the SA process during the data collection phase [16, 50]. Thelwall [50] observed that the process of sampling social media data might result in an unrepresentative sample that distorts findings and limits their generalizability. For example, Twitter users tend to be younger, more affluent, and more urban than a representative sample of the US population [17]. Filtering on websites and keyword searches may also skew samples by excluding critical or more representative data, as algorithms prioritize some posts or terms over others [23], introducing systematic bias into the research process. Studies that focus explicitly on bias in the application of SA tools include, for example: Diaz et al. [16], Goncalves et al. [20]; Iqbal et al. [24], Kiritchenko and Mohammad [26], Kucuktunc and Cambazoglu [29], and Thelwall [50]. These studies considered social bias, race, gender, and age, and test new approaches to bias-aware SA to offset bias in polarity. As a whole, prior research suggests that bias is a systemic problem in SA, emanating from sources such as datasets, training data, other corpora, lexicons, and word embeddings that algorithms draw on to build prediction models [26]. In this way, the range of biases existing within

human-authored texts is unknowingly replicated by systems trained on them [9]. Overall, bias can be introduced into the SA process during its many stages.

2.2.2 Role of Demographics. The studies focusing on social bias are associated with the emerging field of critical algorithm studies. Research in this area explores algorithmic bias, focusing on how systems systematically discriminate against particular individuals and groups [19].

Critical studies have focused on the demographic bias of the tools used for obtaining sentiment scores, usually gender, age, or race. Diaz et al. [16] found age-related bias in word embeddings, which are a type of model in natural language processing. Diaz et al. [16] explored word embeddings and their role in perpetuating age-bias. The authors used a data set containing blog posts from a forum for older bloggers to conduct this inquiry. The data set included 242 sentences, with half containing the term "old" (used as an adjective) in a sentence and the other half containing the replacement word "young" in the same sentence. Diaz et al. used common word embeddings to create semantic analogs for the words "old" and "young" in their sentence templates and then compared output using 10 GloVe word embedding models. Results from Diaz et al.'s experiment indicated that word embeddings encode age-related biases. Expressly, regression results indicated that sentences with "old" adjectives were far more likely to be scored negatively and less likely to be scored positively than sentences with implicitly "young" adjectives.

Thelwall [50] used TripAdvisor hotel and restaurant reviews to assess the influence of gender on the accuracy of SA results. He used a predominantly lexicon-based method, SentiStrength, positing that the tool's lexicon would make it less likely to be influenced by gender-specific terms. To test this hypothesis, Thelwall built five datasets from his pool of hotel and restaurant reviews, grouping reviews according to the number of stars (2, 3, and 4) and topic. For each data set, he identified the gender of review writers, male or female. This permitted assessment of the accuracy of the SentiStrength results separately for male-authored and female-authored reviews. Using mean absolute deviation (MAD) scores, which provide the average of the absolute differences between the original user ratings and SentiStrength, Thelwall determined that the accuracy of the SentiStrength results was significantly lower for males than for females, providing clear evidence of gender bias.

Thelwall [50] also conducted an additional experiment to identify the reason for gender differences in TripAdvisor hotel and restaurant ratings. Toward this end, he compared male-authored reviews to female-authored reviews at each SentiStrength sentiment score level. He observed that SentiStrength significantly underestimated the number of negative reviews for restaurants and hotels and both genders. In addition, female authors tended to express more extreme sentiments (very positive or very negative), whereas males tended to express more mild sentiments. Further analysis indicated that the main reason for gender differences for accuracy was the greater use of patently positive words (e.g. "fabulous" or "amazing") when providing a higher rating and explicitly negative words when giving a lower rating (e.g. "shocking").

Kiritchenko and Mohammad [26] examined sentiment intensity scores, focusing on race and gender. The authors used a self-compiled benchmark dataset ("The Equity Evaluation Corpus") constructed to examine bias in SA tools. They tested the dataset on 219 ML-based SA systems. Researchers analyzed emotion intensity regression to determine race and gender bias for four emotions (anger, fear, joy, and sadness). Overall, 75% to 86% of the systems consistently scored sentences of one gender higher than another, particularly in sentences containing emotions of anger or joy. The race bias analysis produced similarly discrepant results. Most systems assigned higher scores to sentences with African-American names on the task of anger, fear, and sadness, reinforcing prior work, which indicates that African Americans are frequently associated with negative emotions. Alternatively, most systems assigned higher scores to sentences that

contained European American names on the task of joy. Goncalves et al. [20] also observed that most tools presented more positive values than negative values. Moreover, several tools obtained only positive results, regardless of the dataset being used. This positivity bias toward polarity existed using the Twitter logs subsets as well as the Web 2.0 subsets. Even extremely negative events, such as an Air France crash was considered positive by four of the tools.

2.2.3 Individual Differences in Linguistic Styles. Human-authored texts reflect stylistic and syntactic differences in the expression of sentiment [50]. These variations in communication approaches can make it difficult for SA tools to detect what is being described and who is expressing a sentiment. For example, meanings may be communicated differently depending on the social position of an author. If these linguistic differences are not accounted for, SA algorithms may generate erroneous information that underrepresents or misrepresents the attitudes and opinions of certain groups. For example, Thelwall [50] pointed to stylistic and substantive syntactic differences that exist between the genders. These affect word choice, the strength of sentiment, and the use of features (e.g., punctuation). For example, in Twitter-based texts, women are more likely than men to use emotion-related terms, such as love, as well as exclamation marks [11, 52]. These differences indicate that SA algorithms may perform differently when evaluating texts created by or dealing with males and females [50].

2.2.4 Solutions for Bias. To a certain extent, these issues can be managed (as can selection bias in traditional content analysis), and existing studies offer suggestions about working around these pitfalls and mitigating bias. Abdul-Mageed and Diab [2] showed how training annotators on issues related to linguistics improved the annotation process. They also showed how different social media genres might be easier to label for subjective and sentiment than others (e.g., newswire data). The reality of the state-of-the-art thus indicates that SA tools are to be implemented cautiously. While several authors have discussed methods to reduce bias in NLP systems [9, 16, 45, 50, 53], there are no definitive solutions to date. Iqbal et al. [24] presented and tested an approach to mitigating bias in lexicon-based tools, which they called Bias-Aware Thresholding (BAT). Iqbal et al. [24] assessed BAT in relation to two popular lexicon-based tools, AFINN and Sentistrength on seven datasets, incorporating a cross-section of UGC from Twitter, BBC forums, movie review sites, and other online sources. AFINN and Sentistrength are two SA tools that contain words commonly used in UGC (e.g., slang). They obtained quality results in the evaluation – however, the authors recognized there is room for improvement. Ribeiro et al. [41] provide a comparative analysis of research-based SA tools. Although they examine some paid software packages, e.g., LIWC and SentiStrength, they do not inspect commercial API-based SA tools. As we argued in the introduction, such tools are important for software engineers, as they are less dependent on any specific programming languages or environment than research-based libraries or modules.

2.3 Algorithmic Bias in Engineering Systems and the Need for Scrutiny

An adjacent but growing area of interest is the study of algorithmic biases – defined here as systematically categorizing data toward a specific class of sentiment [27]. This risk is particularly alarming for black-box ML models that do not enable the researchers or practitioners access to the detailed decision-making process of the algorithm, thus making direct feature importance analyses and debugging impossible [21]. The SA tools we analyze exemplify black-box models that may exhibit algorithmic bias, as the details have not been made available on how they were trained, how they exactly make the predictions, and how to query the features participating in the algorithm's decision-making process [27]. The risk of bias in API-based SA tools is exacerbated by the widespread use of APIs to engineer interconnected systems [48]. Engineers and scientists increasingly use third-party services, such as SA tools, to score and label comments for applications and research purposes. For these stakeholders, the tools are essentially black-box algorithms, as there is

no way for developers or scientists to directly gauge the reliability of the information the API returns. The need for evaluating the algorithmic bias of such closed systems has been widely noted in academic debate on algorithmic ethics [16]. Thus, investigating bias in sentiment scores is a concern for *both* the validity of the scores (i.e., is the information correct) as well as that of ethics (i.e., can the results be applied in good faith) when applying black-box sentiment tools. Thus, the risk of bias in SA tools is notable and should be investigated, as we do in this study. Here, we investigate the agreement and bias of the popular SA tools, addressing the call for more empirical studies on algorithmic bias in engineering systems [16].

3 METHODOLOGY

3.1 Choice of Sentiment Analysis Tools

There are many SA tools published by researchers based on different techniques and modes of analysis. These tools include, for example, SentiWordNet [6, 8, 15, 20, 24, 39]; Sentistrength [20, 24, 50]; Lydia [7, 54]; PANAs-t [20]; Emoticons [20]; SASA [20]; SenticNet [20]; Happiness Index [20]; LIWC [20]; and AFINN [24]. In addition to research tools, commercial tools are becoming increasingly common in the practice of software development, as they afford easy access to API endpoints for automatic labeling of social media and other online content. Therefore, this study focuses on evaluating those tools, stemming from the fact that while most research tools have been evaluated in previous research, commercial tools have been largely overlooked.

Following this decision, we compare four SA tools from major providers of cloud solutions. Microsoft Azure (MS): Text Analytics version 3.0[4], IBM Watson (IBM): Natural Language Understanding version 2020-08-01[5], Google Cloud (Google): Natural Language version 1[6], and Amazon Web Services (AWS): Comprehend version 2[7] provide Machine Learning as a Service (MLaaS) that covers most infrastructure issues, such as data pre-processing, model training, and model evaluation with further prediction through service APIs. We focus on these four platforms, as they are from the dominant players in providing MLaaS. All four platforms provide software as a service (SaaS), and the facility of integrating them with the ML ecosystems of the respective technology companies makes adopting these tools likely for commercial organizations. Given this, it is critical to identify possible sentiment bias in the SA tools.

3.2 Description of the Tools' Features

IBM and Google output a sentiment score between -1.0 and 1.0. Scores close to 1.0 indicate positive sentiment, and scores close to -1.0 indicate negative sentiment. Additionally, IBM provides the sentiment class (positive, negative, and neutral). MS and AWS return the most likely sentiment and the scores for each sentiment class (positive, negative, neutral, and mixed). For these two tools, the sum of values for the sentiment class for each content equals one; for example, [negative 0.2, positive 0.7, neutral 0.1] == 1.0. English is the most commonly supported language, followed by Spanish, French, German, Italian, Portuguese, Korean, Chinese, Japanese. MS and AWS do not support Arabic and Dutch, respectively, unlike the other tools. Google, unlike the other tools, supports Thai and Vietnamese. AWS and MS both support Hindi. Only IBM supports Russian. Moreover, each tool has a limited text size for determining sentiment. This limit does not notably affect our analysis, as the average content length in our dataset is 113.87 characters (SD = 190.15), with 95.29% of the content within the accepted limits.

[4]https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/
[5]https://www.ibm.com/cloud/watson-natural-language-understanding
[6]https://cloud.google.com/natural-language
[7]https://us-west-2.console.aws.amazon.com/comprehend/v2/home

MS and AWS provide a batch API call, with MS returning 10 results, and AWS returning 25 results per one API call. IBM and Google only handle one text per API call. For 1,000 comments over 10 iterations, AWS returns the sentiment result quickly. The average processing speed of AWS is 23.09 seconds (SD = 2.22 s). Even though MS handles a batch API call, it is not fast as AWS does: 151.10 seconds (SD = 5.52 s). IBM and Google are relatively fairly slow, with IBM taking on average 11.01 minutes (SD = 0.52 min) and Google 9.58 minutes (SD = 0.85 min). So, when it comes to handling many comments for SA, AWS is the most feasible for larger projects.

3.3 Data Collection

Typically, tools are used to analyze one dataset (i.e., one platform/source) in a study. A handful of studies apply SA tools to more than one dataset to determine the text's sentiment. Most of these studies are lexicon-based [6, 7, 20, 24, 50, 55], some used ML [8, 14]; or use a hybrid approach [15, 37, 39]. Multiple datasets have been applied to multilingual and multimodal analyses; to compare genres/domains; and to determine bias.

Since we could not locate a publicly available dataset for online news that would contain news stories and comments from multiple online platforms, we created a new dataset called the **NEWS** dataset. Our dataset originates from a major online news and media organization's content on multiple platforms for this research. This media organization publishes news content on Facebook (FB), Instagram (IG), Twitter (TW), and YouTube (YT). Using the social media platforms' APIs, with the organization's permission, we collect a total of 657,875 user comments from 18,522 news stories from the platforms that are recently published.

To corroborate the results on another dataset, we employ a publicly available **IMDB** dataset that contains 50,000 movie reviews[8]. The IMDB dataset consists of 25,000 positive reviews and 25,000 negative reviews. Using these values as ground truth, we are able to compute the accuracy of each commercial SA tool to provide implications for software engineers – these results are provided in the practical implications section.

3.4 Data Processing

Data processing involved discarding 14,291 comments with no text content and 764 comments written by the organization's account. We remove links from texts – 4,634 comments contain only links, so these comments are also eliminated. The name tag designating the organization account is removed from comments (e.g., all the TW replies have the name tags to reply to the organization account). We eliminate the timestamps from comments of YT (e.g., @11:15). We consider the comments starting with a name tag designating a user other than the organizational account, as conversations with other accounts, not direct comments on the content itself. Since we want to analyze the sentiment of the comments aligned with the news story, not off-topic comments, we exclude these 24,741 comments containing a name tag.

Additionally, we only consider the comments written in English, as the SA tools all support English. By using two Python libraries for language detection, langdetect[9] and langid[10], 184,151 comments are classified as not written in English, so they are eliminated.

Moreover, HTML tags are replaced with proper text characters (e.g., """ to plain double-quotes) using the Python library html2text[11]. The text in hashtags might have a critical impact on sentiment detection, so we keep the text on hashtags after removing the hash sign ("#"). There are some texts with the under-dash ("_"). The texts surrounding

[8] https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
[9] https://pypi.org/project/langdetect/
[10] https://pypi.org/project/langid/
[11] https://pypi.org/project/html2text/

the under-dash might be meaningful, so "_" is replaced with an empty space. For IMDB, some reviews have a HTML tag for a line change, so the tag is removed from reviews.

After all the pre-processing steps, 429,294 comments remain as prepared data for the study. To create a balanced dataset, we randomly select 15,000 comments for each social media platform from the prepared comments (in total, 60,000 comments). This was considered to be an adequate sample to evaluate the sentiment differences both by the SA tools and the online platforms. The comments originate from 5,880 news stories (FB = 1,281, IG = 503, TW = 3,346, YT = 750). In addition, we randomly sample 7,500 positive and 7,500 negative reviews for IMDB dataset.

3.5 Assignment of Sentiment Labels

Most of the sentiment of sampled comments, corresponding news stories, and reviews were successfully detected by the four SA tools (see Table 1). Regarding failures, IBM returned an error stating that the failed instances did not have enough text, or the language was not supported. Google also returned some errors indicating that the given language was not supported. When MS and AWS returned an error, the reason was that the text size exceeded the limit. So, we were not included in the analysis.

	AWS	Google	IBM	MS
News stories (NEWS)	5,879	5,878	5,876	5,880
Comments (NEWS)	59,988	59,998	57,629	59,992
Movie reviews (IMDB)	14,982	15,000	14,995	15,000

Table 1. Summary of comments, news stories, and reviews successfully classified by the SA tools. All tools are able to label almost all content.

Three tools (MS, AWS, and IBM) return a sentiment class (positive, neutral, negative), so we can use these labels without transforming. However, Google only returns a sentiment score ranging from -1 to 1. To make Google's sentiment scores commensurable for analysis, they were transformed to an equivalent nominal scale (positive, neutral, negative) as follows: 0 = neutral, < 0 = negative, > 0 = positive. Unlike IBM, MS and AWS return an additional sentiment value called *mixed*. Both tools do not share the determination much for the sentiment *mixed*. Especially for movie reviews (the IMDB dataset), MS tends to return *mixed* sentiment more than AWS (See Table 2). As there is no equivalent score in IBM and Google for the sentiment *mixed*, we ignore the mixed comments in the following analysis. This was not an issue for the NEWS dataset, but the IMDB sample now decreased to 2,230 reviews (14.9% of total).

	AWS	MS	Shared	Total (%)
News stories (NEWS)	39	388	14	413 (7.0%)
Comments (NEWS)	2,827	5,479	714	7,592 (12.7%)
Movie reviews (IMDB)	3,076	12,496	2,802	12,770 (85.1%)

Table 2. Content classified as *mixed* by MS and AWS. The mixed content items were excluded from the analysis because this category was not provided by IBM and Google.

4 RQ1: HOW WELL DO DIFFERENT SA TOOLS AGREE ON THE EXACT SAME CONTENT?

4.1 Approach

To address RQ1, we calculate the inter-rater agreement score among the tools using Fleiss' kappa (κ), a statistical measure for assessing the reliability of agreement between a fixed number of raters (in our case, each tool is considered a rater, which means we have four raters) when assigning categorical ratings to N items. Fleiss' κ is based on an N by k observation table or matrix in which the elements n_{ij} indicate the number of raters who assigned the i-th case in the j-th category. The agreement attained in excess of chance $P_o - P_e$, normalized by the maximum agreement attainable above chance $1 - P_e$, defines the κ statistic, as shown in Equation 1:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

We calculated the κ between the tools, both overall and by platform. The metric considers the degree of agreement over that which would be expected by chance. The value of κ can range from 0 to 1. Despite there being no universal way to interpret κ, the values proposed by Landis and Koch [30] can be used for primary interpretation, as we do in the next section. The interpretation of the agreement is as follows: slight (0-0.20), fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80), and almost perfect (>0.80). Confidence intervals (CI) are reported at 95%.

4.2 Results

For the NEWS data, the overall agreement among all tested SA tools is fair (κ = 0.48, CI = 0.473, 0.486, p < .001). All pairwise agreements are also in the range of fair agreement (0.41–0.60), with some reaching a moderate agreement. Still, the pairwise agreement scores do show some variation (see Table 3). The sentiment labels between IBM and MS have the highest agreement (κ = 0.547, CI = 0.541, 0.553, p < .001) among all pairs, indicating that IBM and MS more often agree with each other than any other pair. A relatively high agreement score is also obtained by IBM and AWS (κ = 0.519, CI = 0.513, 0.525, p < .001) and MS and AWS (κ = 0.507, CI = 0.501, 0.513, p < .001). These results indicate that IBM has a tendency of agreeing with the other tools more often. In contrast, Google and AWS agree the least with other tools (κ = 0.396, CI = 0.389, 0.403, p < .001).

	AWS	Google	IBM	MS
AWS	1.00	0.40 (0.79)	0.52 (0.77)	0.51 (0.76)
Google	0.40 (0.79)	1.00	0.44 (0.86)	0.44 (0.85)
IBM	0.52 (0.77)	0.44 (0.86)	1.00	0.55 (0.91)
MS	0.51 (0.76)	0.44 (0.85)	0.55 (0.91)	1.0

Table 3. Pairwise agreement (Fleiss' κ) among the SA tools. IMDB values in parentheses, NEWS values not in parentheses. While the agreement of the tools tends to be fair for the NEWS dataset, for the movie reviews it is substantial or almost perfect in all cases.

There are also platform-specific differences in the NEWS dataset (see Table 4). For FB, IBM and MS have the highest agreement (κ = 0.539, CI = 0.527, 0.551, p < .001) that is higher than the overall agreement. In turn, the lowest score found between Google and AWS (κ = 0.377, CI = 0.363, 0.390, p < .001) is smaller than the overall minimum score. Moreover, the agreement scores of all pairs with Google are less than κ = 0.43, indicating that Google is highly likely to disagree with other tools when classifying the sentiment of FB news stories and comments.

Comparing the tools' average agreement scores by platform, we observe that IG has the highest agreement (κ = 0.528, CI = 0.516, 0.541, p < .001). This is 13.9% higher than the average agreement score of all the platforms. FB's agreement

	Fleiss' κ	SE	z	p	Cl lower	Cl upper
FB	0.468	0.006	72.904	<.001	0.455	0.480
IG	0.528	0.006	81.661	<.001	0.516	0.541
TW	0.428	0.006	68.013	<.001	0.416	0.441
YT	0.428	0.008	57.011	<.001	0.414	0.443

Table 4. Platform-specific agreements in the NEWS dataset. The agreement is consistently in the fair range regardless of the platform.

score (κ = 0.468, CI = 0.455, 0.480, p < .001) is also higher-than-average (by 1.0%). In contrast, TW (-7.5%, k = 0.428, CI = 0.416, 0.441, p < .001) and YT (-7.5%, κ = 0.428, CI = 0.414, 0.443, p < .001) have lower-than-average agreement. For all tools, negative comments generally outnumber the positive comments across all the social media platforms (see Table 5). There are 2.9 times more comments labeled as negative than positive. Negative sentiment appears 5.7 times more than positive sentiment in YT. In contrast, FB and TW have a similar proportion of negatively and positively labeled comments, indicating a more balanced sentiment tendency for these platforms.

We observe substantial differences when comparing the results obtained from the NEWS dataset to those from the IMDB dataset. The total agreement among the four tools regarding IMDB reviews is substantially higher (κ = 0.82, CI = 0.796, 0.845, p < .001). The same can be observed in the number of instances where all tools agree.

	negative	neutral	positive	total (N)
FB	66.4%	8.2%	25.4%	6,178
IG	63.4%	6.0%	30.6%	6,534
TW	68.6%	7.3%	24.1%	5,726
YT	80.3%	5.7%	14.0%	6,233
Overall (NEWS)	69.6%	6.8%	23.6%	24,671

Table 5. Summary of sentiment labels for user comments where all tools agreed.

The number of instances in the NEWS dataset where all tools agree is 25,853 (46.6%). The number of instances in the IMDB dataset where all tools agree is 1,902 (85.4%). The number of total instances (NEWS + IMDB) where all tools agree is 27,755 (48.1%).

We also examined the content types within the NEWS dataset, and found that news stories had a substantially lower agreement (κ = 0.20) than news comments (κ = 0.49). This is interesting because the news stories actually include more content (M = 237.5 characters, SD = 294.8) than the comments (M = 104.3 characters, SD = 114.7).

Addressing RQ1, the results are mixed. The SA tools predominantly disagree about the labels in the NEWS dataset but agree about labels in the IMDB dataset. Since the IMDB dataset is publicly available, it is possible that some or all providers of the tools are using it to train their models. The other possible explanation is that movie reviews are less ambiguous content type than news stories and their comments. Nonetheless, the sharp difference in the agreement scores indicates that the agreement of commercial SA tools depends on the dataset.

5 RQ2: ARE SA TOOLS SYSTEMATICALLY BIASED IN ASSIGNING SENTIMENT LABELS FOR THE EXACT SAME CONTENT?

5.1 Approach

To address RQ2, we first explore the frequencies of the output labels provided by the SA tools. Second, we evaluate the bias of the tools by measuring the tendency of each SA tool to detect each sentiment label relative to the average of all tools. For this, we calculate the relative risk ratio (RR) for each tool t and for each sentiment label s using Equation 2:

$$RR(s,t) = \frac{N(s,t)}{\frac{1}{M}\sum_{i=1}^{M} s_i}$$

where $N(.)$ is the number of content items for sentiment label s and tool t. M is the total number of content items across all the tools for sentiment labels s_i, where $i = 1, 2, 3...M$. The denominator in Equation 2 thus describes the average number of content items across all the tools for the given sentiment label s. A higher RR indicates that a tool tends to indicate a specific sentiment label more than the average of all the tools.

5.2 Results

5.2.1 Label Frequencies. Table 6 shows the distribution of sentiment from each SA tool across the social media platforms after these transformations. It can be seen that MS tends to indicate *mixed* sentiment relatively more than AWS. In turn, AWS is likely to classify the sentiment as neutral than the other SA tools. The news stories across platforms are mostly negatively classified by the SA tools, and the comments also are indicated as negative sentiment. The sentiment distribution of IMDB reviews shows the tendency of each SA tool (see Table 7). MS classifies the sentiment mostly as *mixed*. A high prevalence of the mixed class indicates that the tool is uncertain and cannot assign a definite label for the content. While this class can be useful for communicating model uncertainty to users of SA tools, its large prevalence for AWS and IBM hurts the accuracy of these tools, as we discuss in the implications. The sentiment distribution of IMDB reviews from the SA tools (see Table 7) indicates the dataset might be used by IBM as a part of the base dataset since none of the reviews is classified as neutral sentiment. In total, 84.7% of the reviews are correctly determined their sentiment by IBM. MS also does not have neutrally classified reviews, but most reviews are determined as mixed sentiment. In turn, AWS tends to label most content in the NEWS dataset as neutral, but only some content as neutral in the IMDB dataset. Among the sampled IMDB reviews, only 1,902 reviews gain unanimity for sentiment determination from all SA tools, and among them, 96.9% of reviews are classified as the same sentiment with the ground truth.

5.2.2 Relative Risk Ratios. The results in Figure 1a indicate that AWS tends to classify sentiment as neutral 1.56 times more than the average in the NEWS dataset (Figure 1a. As AWS has a strong tendency to indicate sentiment as neutral, it has the lowest ratio for both positive (RR = 0.73) and negative (RR = 0.86) sentiment labels. The same pattern of neutrality bias for AWS can be observed in the IMDB dataset (see Figure 1b Therefore, AWS can be characterized as "undecided" or "ambivalent" relative to other tools concerning the NEWS dataset. In the NEWS dataset, IBM and MS can be seen as the most "bias free," as their output labels for all classes correspond quite closely to average values (see Figure 1a. The labeling pattern between IBM and MS also remains close to identical for the IMDB dataset (see Figure 1b), which implies that these models behave in a similar way across different datasets. The pattern is different in the IMDB dataset, which IBM and MS interpret to contain only positive and negative cases. This interpretation is correct,

		News stories				Comments			
		mixed	negative	neutral	positive	mixed	negative	neutral	positive
FB	AWS	0.3	18.9	78.9	1.9	4.2	44.6	33.5	17.7
	Google	0.0	68.8	15.9	15.3	0.0	60.2	11.7	28.0
	IBM	0.0	45.7	39.7	14.6	0.0	53.6	22.2	24.2
	MS	3.2	65.2	25.4	6.2	7.3	44.0	23.8	24.8
IG	AWS	1.6	20.9	75.0	2.6	5.3	44.0	28.4	22.3
	Google	0.0	55.7	27.3	17.0	0.0	56.5	11.5	32.0
	IBM	0.0	54.9	14.3	30.8	0.0	54.3	16.5	29.2
	MS	19.1	54.3	14.3	12.3	9.8	43.0	20.2	27.0
TW	AWS	0.5	21.7	75.4	2.4	4.4	44.6	33.9	17.2
	Google	0.0	64.1	16.9	19.0	0.0	60.6	11.7	27.7
	IBM	0.0	39.7	49.0	11.3	0.0	55.8	20.3	23.9
	MS	2.2	54.2	33.1	10.4	7.7	45.3	24.3	22.7
YT	AWS	1.5	27.5	71.1	0.0	5.0	56.3	28.1	10.6
	Google	0.0	84.9	11.3	3.7	0.0	71.3	11.2	17.5
	IBM	0.0	73.4	12.3	14.3	0.0	63.8	17.7	18.5
	MS	23.7	68.1	4.7	3.5	11.8	50.3	19.8	18.0

Table 6. Sentiment distribution by the SA tools across the social medial platforms in percentage.

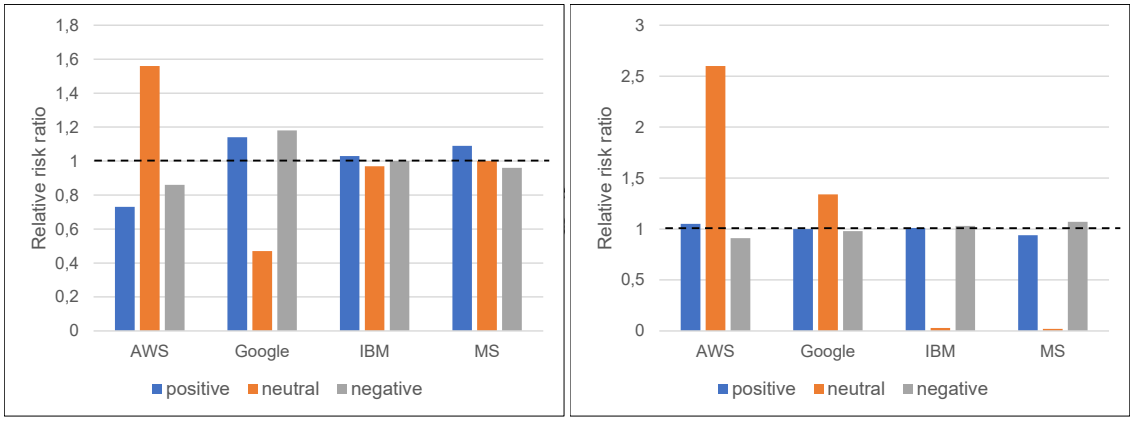
	AWS	Google	IBM	MS
mixed	20.5% (n = 3,076)	0.0% (n = 0)	0.0% (n = 0)	83.3% (n = 12,476)
negative	36.4% (n = 5,459)	47.4% (n = 7,102)	56.8% (n = 8,503)	11.9% (n = 1,778)
neutral	8.6% (n = 1,287)	14.8% (n = 2,220)	0.0% (n = 2)	0.0% (n = 1)
positive	34.4% (n = 5,156)	37.8% (n = 5,656)	43.2% (n = 6,473)	4.8% (n = 723)

Table 7. Sentiment distribution of IMDB reviews by the SA tools. An interesting observation here is the mixed class, which can be seen as a way to quantify uncertainty in the SA tools' predictions.

as the IMDB dataset only has positive and negative cases (we will discuss the accuracy of the tools in the practical implications section).

Google's behavior is interesting – for the NEWS data, Google is 1.18 and 1.14 times more likely to indicate negative and positive sentiment more, respectively, than the average across the tools. This characterizes Google's interpretation of the content as "bi-polar" relative to other tools. However, this behavior changes with the dataset – for the IMDB data, Google is now more inclined to choose the neutral class (see Figure 1b), effectively displaying a labeling pattern similar to AWS.

Addressing RQ2, the results indicate that AWS exhibited neutrality bias in both datasets. Google exhibited bi-polarity bias in the first dataset but neutrality bias in the second dataset. IBM and MS exhibited no apparent bias in the first dataset but had bi-polarity bias in the second dataset. The different biases in the two datasets indicate that observed sentiment biases depend on the dataset. The ground truth sentiments were known for the IMDB but not for the NEWS dataset. However, the different outcomes imply that the distribution of sentiments in the NEWS dataset differs from that in the IMDB dataset. The reader should also note that the observed biases were determined by examining RRs relative to the average label values provided by all tools, not against known ground truth values. This might have affected the results regarding the IMDB dataset, namely in that IBM and MS are not biased in their bi-polar interpretation but actually accurate, as the dataset only contains positive and negative samples. This is further elaborated in the discussion section.



(a) (b)

Fig. 1. Relative risk ratios for NEWS dataset (a) and IMDB dataset (b). Dashed lines indicate the average of all tools. Values higher than that indicate a tendency to be more inclined to a class (e.g., AWS in being more inclined to neutrality than the other tools), whereas values lower than the average indicate a tendency to be less inclined to a class.

6 RQ3: WHY DO THE TOOLS DISAGREE?

To address RQ3, we investigated why the tools agree or disagree. For this, we conducted binary-class experiments with various ML algorithms commonly used for text analysis (Logistic Regression, Support Vector Machines, Naïve Bayes, and LightGBM). The algorithms predicted two classes: Agreement (the subset of the data in which all tools agree about the sentiment label) and Non-agreement (the subset of the data in which only two tools agree). Given three sentiment labels and four tools, two tools agreeing represents the minimum possible agreement). As such, the models predicted if the tools would agree or disagree on the sentiment label for a given content.

We did not use word embeddings in our analyses, as we wanted the features to be interpretative. The features included Bag-of-Words (BOW) [38], Term Frequency – Inverted Document Frequency (TF*IDF) [38], and LIWC (Linguistic Inquiry and Word Count) [49]. LIWC processes each text sample, locates individual words, compares them with its built-in dictionary, counts that word, and increments as a straight count and then applies that count to a simple percentage function (% of the complete document) or a variable-based scale function (based on dedicated algorithms) for psychometric, psycholinguistic, or other human-related measures.

The LightGBM algorithm with all features provided the best accuracy (ACC = 0.709). We then proceeded to infer the feature importance values using the SHAP framework [32] that outputs how much each feature contributes to the predictions. Results in Figure 2 show that the LIWC features are most impactful for the predictions. Out of the 20 most impactful features, 18 are LIWC features. LIWC is a human-curated lexicon that assigns social meanings to words observed in text samples. Apart from these, two particular words, “good” and “love” are considered impactful by the model when explaining why the tools agree or disagree on a sample.

Addressing RQ3, the results indicate that, apart from a few TF-IDF features such as the use of positive words ‘good’ and ‘love’, manually curated psycholinguistic features (i.e., LIWC) are stronger predictors for whether the tools disagree or not. The most impactful features are affective expression (‘affect’), emotional tone, the use of adjectives (‘adj’) and swear words (‘swear’), and dictionary coverage (Dic). Interestingly, these features are more impactful than the positive and negative classes in the LIWC dictionary, as these two classes rank only sixth and seventh most important,

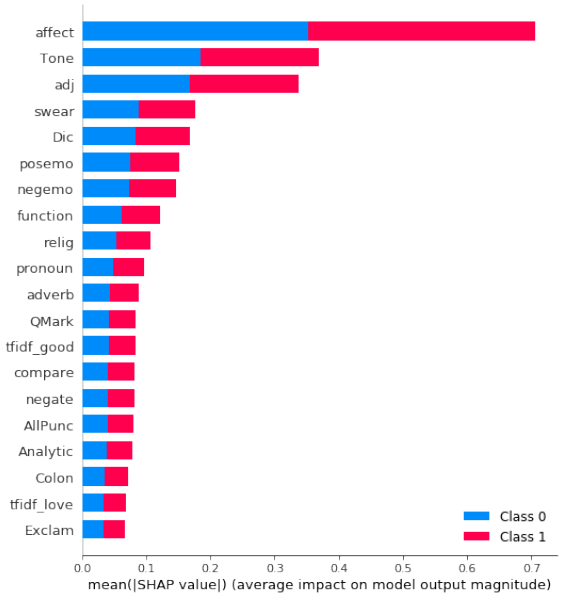


Fig. 2. TOP 20 most impactful features for predicting if the SA tools agree or disagree on a sample of text. Apart from the words 'good' and 'love', all of the most impactful features are from LIWC. The most impactful features relate to affect (affect), emotional tone (Tone), and the use of adjectives (adj) and swearing (swear). Interestingly, these features are more impactful than the positive and negative features in the LIWC dictionary. Class 0 indicates agreement and Class 1 disagreement.

respectively. The 'affect' class measures affective or emotional processes, with several hundreds of manually curated patterns [35]. The emotional tone class ('Tone') measures the tone of written messages – the higher the score, the more positive the tone [49]. It is, therefore, logical that these classes predict sentiment disagreement – in other words, different tools interpret affective and emotional expressions in different ways: some may pick a certain expression while others do not; some may consider the expression positive in a given context while others do not. The fifth most impactful class – dictionary word count ('Dic') – indicates that uncommon words (i.e., those not available in a broad dictionary such as LIWC) play a role in sentiment disagreement.

7 DISCUSSION

7.1 Main Contribution

Previous research has shown that people, such as crowdworkers assigning sentiment labels, tend to disagree on the label [5, 33, 34, 42]. Ribeiro et al. [41] showed that research-based SA tools exhibit different biases, but as far as we know, no other study has examined the agreement and bias in commercial API-based SA tools. Because software engineers are increasingly relying on commercial API tools due to their simplicity and support relative to research-based tools, it is vital to investigate the agreement and bias in these tools. Such study is even more important because, in the software engineering and managerial fields, there is often an expectation of objectivity from numbers provided by opaque algorithms [44] – as our results show, this is not the typical case for API-based SA tools.

The results indicate that tools' disagreement depends on the dataset. This can be due to the difficulty of different datasets or because a dataset was already included in the tools' training data. The fact that both agreement and accuracy

were substantially higher on the publicly available IMDB dataset than on the private NEWS dataset implies that this dataset could have been used in the SA tools' training process. However, it is impossible to know that because of the tools' lack of transparency. Similar to the agreement, also biases were for the two datasets. The patterns were similar between MS and IBM for both datasets, meaning these tools had a similar bias for each dataset. AWS's neutrality bias appeared for both datasets, implying this pattern of bias is a more stable trait in AWS's labeling behavior. Google changed from bi-polar bias in the NEWS dataset to neutrality bias in the IMDB dataset, indicating sporadic behavior, especially because the IMDB dataset did not have a neutral class (based on ground truth labels).

Concerning the NEWS dataset, the modest correlation of sentiment labels given by the four major sentiment tools indicates that the SA tools struggle to agree on the sentiment of online news comments. Concerning the IMDB dataset, the high agreement indicates that the sentiment in movie reviews is easier to interpret by the tools (or that this data was already seen by the SA tool). In any case, our findings demonstrate that SA tools are not purely objective, but each tool tends to have a distinct pattern of labeling (with MS and IBM being very similar to each other).

The classification results on the agreed and disagreed cases indicate that psycholinguistic features can contribute to explaining tool (dis)agreement. Particularly impactful features relate to expressions of emotion and affect and the use of adjectives and curse words. Words in these categories can have a sentimentally different meaning depending on the context – for example, cursing can sometimes be a sign of affection. Affective expressions could be used for undermining others in a social media conversation, and so on. The feature importance analysis shows that the SA tools interpret certain categories of words differently, resulting in disagreement over whether a content item is positive, negative, or neutral. The influence of context is particularly distinctive for social media conversations, in which users exhibit many traits that make the language difficult to classify, including spelling mistakes, frequent use of irony and other forms of humor (e.g., dark jokes), slang words that may not be included in a tool's dictionary, and so on. According to what we term the *chaos theory of social media content*, social media content is simply too random, chaotic, and mixed for any number of tools to agree on the sentiment of a large body of content. Each tool may address different aspects of language, thereby distinctly interpreting the content. If this theory holds, then seeking an agreement might turn out to be futile, as there is no definitive agreement at all.

Developing this thought further, in the absence of absolute agreement, researchers should instead focus their efforts on questions such as how to communicate disagreement and biases to end-users to make informed decisions instead of believing in the one sentiment score or label being provided as one single truth. Effort to this end can focus on tasks such as uncertainty quantification – i.e., how to communicate model uncertainty, perhaps using multiple tool disagreement as a proxy metric when lacking direct access to the model, as happens with API-based tools. Uncertainty quantification is a nascent field in the ML community [1], and SA tools provide a highly compatible domain of inquiry due to their prominence and inherent uncertainty (i.e., disagreement) in their outputs.

7.2 Practical Implications for Mitigating Bias in Engineering Interactive Systems

The fact that the SA tools often disagree (at least on social media NEWS data) has profound implications for implementing SA tools in real systems, as it raises profound questions for engineering interactive systems, such as *Can we rely on the sentiment scores given by SA tools at all? How should we communicate the sentiment labels to end-users of downstream applications?*

From an engineering point of view, various unknowns affect how commercial providers use to train their ML models, including class (im)balance, guidance given to the human annotators, neural network architectures, choice of hyperparameter values for the algorithm, and the use of (majority) voting techniques to assign the final label for

the samples. As these details are not made public by the SA tool providers, they cannot be scrutinized or improved upon by engineers implementing the tools. Therefore, as it currently stands, due to their black-box nature, there is no definitive solution for bias elimination when using commercial SA tools. Commonly applied approaches are typically not applicable to black-box models because researchers cannot retrain these models. As such, the existence of bias in SA systems is widely acknowledged, but it has not been examined for black-box models, such as the commercial API-based SA tools. Nevertheless, commercial SA tools are widely used by software engineers, often with little consideration of bias.

Moreover, we can provide recommendations for different stakeholder groups:

(1) **For software engineers:** (a) Use several tools and compare their agreements. Choose either one tool (the least biased for your task) or the majority vote / consensus of several tools when assigning final labels. Majority voting is typically used for solving disagreement among human annotators [43]. Although this increases design complexity, this type of corroboration is needed to portray online sentiments in a truthful manner. (b) Keep in mind the online SA tool tends to be upgraded at a certain time and the prediction result could change over time. (c) When using multiple tools, illustrate their agreement / disagreement to end-users in order to increase the transparency of the developed sentiment analysis application.

(2) **For SA tool providers:** Make methods of training, training datasets used, and algorithms used publicly available. At the very least, descriptions of the data used for training should be given so that the organizations and researchers applying the tools could judge the reliability and potentially address deficiencies by training their own models.

(3) **For end-users of SA tools:** Question the tools' reliability and request for more information on how the sentiment labels are assigned and what are the weaknesses of the used approaches.

Finally, to provide further recommendations for engineers, we computed the accuracy (ACC) of the tools using the IMDB dataset that had ground truth values. The results (see Table 8) show that IBM outperforms other tools (ACC = 84.6%). Because the models' training data is not made publicly available, it is unknown if IBM would be trained on samples from the IMDB dataset – however, the accuracy is not perfect or even above 90%, which implies that IBM at least does not perfectly know the IMDB dataset. The low performance of MS (ACC = 14.9%) is interesting and can partially be explained by MS's large share of undetermined samples from the mixed class (see Table 7). Because the ground truth has no "uncertain" cases, AWS and MS that quantify model uncertainty with their mixed class are penalized. In conclusion, out of the tested tools, IBM seems to be the best choice based on its lack of bias in the NEWS dataset and its best performance on the IMDB dataset.

	Positive	Negative	Overall	Accuracy
AWS	4,607	4,892	9,499	63.3%
Google	5,395	6,372	11,767	78.4%
IBM	5,841	6,861	12,702	**84.6%**
MS	701	1,542	2,243	14.9%

Table 8. The accuracy of each tool for 15,000 sampled reviews from IMDB dataset. Best result bolded.

7.3 Limitations and Future Work

Our work involves some limitations. First, our definition of "bias" is provided without a known ground-truth sentiment of the comments. We provide the following justification for this choice: since the labeled text (comments and news stories) are constant, the patterns that we show indicate bias existing in one tool relative to the other tools, regardless of the direction. Hence, all the reported biases (neutrality bias by AWS and bi-polarity bias by other tools in different datasets) were determined by examining risk ratios relative to the average labels provided by all tools, not against known ground truth values. This might have affected the results regarding the second dataset, namely in that IBM and MS are not biased in their bi-polar interpretation but actually accurate, given that the dataset only contains positive and negative samples.

One factor hindering the validity may be changes taking place in the SA tools. Since the algorithms used by the commercial companies are not public and can change at any time, we conducted a repeated test to evaluate the consistency of results over time. The original sentiment labeling was conducted in December 2019. A repeated test with the content (from the NEWS dataset) was conducted in April 2021. The results showed 70.3% overall similarity, 87.8% similarity on the original negative class (i.e., this proportion of negative content remained negative in the second round), 53.7% similarity on the original neutral 63.8% similarity on the original positive class. The most consistent tool was IBM (96.1% consistency rate), followed by AWS (68.8%), Google (62.2%), and MS (54.2%). The low consistency by MS is explained by the fact that MS has added a new sentiment label for the content of which it is uncertain. While the results are cross-sectional, we surmise that the main finding – disagreement among the tools – is likely to persist over time because this is caused by the companies' idiosyncratic algorithm design and data harvesting.

The fact that people disagree on the sentiment of social media comments and other content might be among the root causes of why the SA tools disagree as well. Therefore, the SA tools have been trained on data that possibly inherit human disagreement (and bias) rather than addressing it. Thus, uncertainty for a given content's sentiment may be integrated with the very decision-making process of the SA tool via techniques such as majority voting that smooths disagreements during the labeling process [43]. Therefore, future research could investigate specific instances where human raters and SA tools disagree – such interpretative studies could help better understand which disagreements are solvable and under what conditions.

The fallacy of perfection regarding the API tools in the software industry should be addressed with more transparency about cases where the tools disagree, as well as quantifying uncertainty in the labeling process and displaying these measures of uncertainty to end-users, so that they can make informed decisions. Therefore, future research should focus on user interfaces and interaction techniques for communicating SA tools' disagreement and uncertainty to end-users. Some of the tools, namely MS and AWS, incorporate a mixed class that is a helpful direction to assign uncertainty when the tool is undetermined. According to our results, these cases constitute roughly one-fourth of the sampled content (based on AWS and MS), which is a sizeable chunk of data.

Finally, future research could compare the bias and agreement of research-driven SA tools (e.g., those investigated by Ribeiro et al. [41]) relative to commercial SA tools, as such a comparison could yield interesting insights on the state-of-the-art dynamics between industry and research community. Our study focused on commercial tools, as these are easy to deploy by both researchers and practitioners and, therefore, present a realistic solution for many software engineers to create applications that analyze social media users.

8 CONCLUSION

The tested tools agree little on social media content, but they agree more on movie reviews, which might be because the movie reviews were publicly available and could thus have been included in the tools' training process. AWS exhibited neutrality bias in both datasets. Google exhibited bi-polarity bias in the first dataset, but neutrality bias in the second dataset. IBM and MS exhibited no apparent bias in the first dataset, but had bi-polarity bias in the second dataset. Psycholinguistic features, such as affect, emotional tone, and adjectives, contribute to explaining why tools agree or disagree, likely because these features represent emotional ambiguity in online contexts. Our results imply that engineers should be cautious when applying sentiment analysis tools to avoid presenting fallacious certainty to end-users of sentiment analysis applications. Out of the tested tools, IBM seems to be currently the best choice based on its lack of bias in the NEWS dataset and its best performance on the IMDB dataset. However, as tools continuously evolve, it is important to repeat tests before implementation or use several tools and display their (dis)agreement to end-users for more transparent sentiment analysis.

REFERENCES

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, and U. Rajendra Acharya. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. (2021). Publisher: Elsevier.

[2] Muhammad Abdul-Mageed and Mona T. Diab. 2012. AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis. In *LREC*, Vol. 515. 3907–3914.

[3] Kholoud Khalil Aldous, Jisun An, and Bernard J Jansen. 2019. View, Like, Comment, Post: Analyzing User Engagement by Topic at 4 Levels Across 5 Social Media Platforms for 53 News Organizations. In *Proceedings of the International AAAI Conference on Web and Social Media*.

[4] D. Alessia, Fernando Ferri, Patrizia Grifoni, and Tiziana Guzzo. 2015. Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications* 125, 3 (2015).

[5] Omar Alonso, Catherine C. Marshall, and Marc Najork. 2015. Debugging a Crowdsourced Task with Low Inter-Rater Agreement. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries* (2015) *(JCDL '15)*. ACM, New York, NY, USA, 101–110. https://doi.org/10.1145/2756406.2757741

[6] Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 100–107.

[7] Mikhail Bautin, Lohit Vijayarenu, and Steven Skiena. 2008. International sentiment analysis for news and blogs. In *Proceedings of the International AAAI Conference on Web and Social Media*.

[8] Adam Bermingham and Alan F. Smeaton. 2010. Classifying sentiment in microblogs: is brevity an advantage?. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. 1833–1836.

[9] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*. 4349–4357.

[10] Luca Buccoliero, Elena Bellio, Giulia Crestini, and Alessandra Arkoudas. 2020. Twitter and politics: Evidence from the US presidential elections 2016. *Journal of Marketing Communications* 26, 1 (2020), 88–114.

[11] John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. 1301–1309.

[12] Erik Cambria. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems* 31, 2 (2016), 102–107.

[13] Andrea Ceron, Luigi Curini, Stefano M Iacus, and Giuseppe Porro. 2014. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New media & society* 16, 2 (2014), 340–358.

[14] Dushyant Singh Chauhan, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Context-aware interactive attention for multi-modal sentiment and emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5651–5661.

[15] Kerstin Denecke. 2008. Using sentiwordnet for multilingual sentiment analysis. In *2008 IEEE 24th international conference on data engineering workshop*. IEEE, 507–512.

[16] Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 412.

[17] Maeve Duggan and Joanna Brenner. 2013. *The demographics of social media users, 2012*. Vol. 14. Pew Research Center's Internet & American Life Project Washington, DC.

[18] Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining.. In *LREC*, Vol. 6. Citeseer, 417–422.

[19] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)* 14, 3 (1996), 330–347. Publisher: ACM New York, NY, USA.

[20] Pollyanna Gonçalves, Matheus Araújo, Fabrício Benevenuto, and Meeyoung Cha. 2013. Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks*. 27–38.

[21] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (2016). ACM, 2125–2126.

[22] Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

[23] Lucas D. Introna and Helen Nissenbaum. 2000. Shaping the Web: Why the politics of search engines matters. *The information society* 16, 3 (2000), 169–185. Publisher: Taylor & Francis.

[24] Mohsin Iqbal, Asim Karim, and Faisal Kamiran. 2015. Bias-aware lexicon-based sentiment analysis. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*. 845–850.

[25] Soon-Gyo Jung, Jisun An, Haewoon Kwak, Joni Salminen, and B. J. Jansen. 2018. Assessing the Accuracy of Four Popular Face Recognition Tools for Inferring Gender, Age, and Race. In *Proceedings of International AAAI Conference on Web and Social Media (ICWSM 2018)* (San Francisco, California, USA, 2018-06-25). https://ojs.aaai.org/index.php/ICWSM/article/view/15058

[26] Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508* (2018).

[27] Sotiris Kotsiantis, Dimitris Kanellopoulos, and Panayiotis Pintelas. 2006. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering* 30, 1 (2006), 25–36.

[28] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg!. In *Fifth International AAAI conference on weblogs and social media*.

[29] Onur Kucuktunc, B. Barla Cambazoglu, Ingmar Weber, and Hakan Ferhatosmanoglu. 2012. A large-scale sentiment analysis for Yahoo! answers. In *Proceedings of the fifth ACM international conference on Web search and data mining*. 633–642.

[30] J. Richard Landis and Gary G. Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* (1977), 363–374.

[31] Shiyue Li, Zixuan Liu, and Yanling Li. 2020. Temporal and spatial evolution of online public sentiment on emergencies. *Information Processing Management* 57, 2 (2020), 102177. https://doi.org/10.1016/j.ipm.2019.102177

[32] Scott Lundberg. 2018. shap: A unified approach to explain the output of any machine learning model. https://github.com/slundberg/shap original-date: 2016-11-22T19:17:08Z.

[33] Isa Maks and Piek Vossen. 2012. A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems* 53, 4 (2012), 680–688.

[34] Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual Twitter sentiment classification: The role of human annotators. *PloS one* 11, 5 (2016), e0155036.

[35] Alexander Osherenko and Elisabeth André. 2007. Lexical Affect Sensing: Are Affect Dictionaries Necessary to Analyze Affect? In *Affective Computing and Intelligent Interaction*, Ana C. R. Paiva, Rui Prada, and Rosalind W. Picard (Eds.). Vol. 4738. Springer Berlin Heidelberg, 230–241. https://doi.org/10.1007/978-3-540-74889-2_21 ISSN: 0302-9743, 1611-3349 Series Title: Lecture Notes in Computer Science.

[36] Bo Pang and Lillian Lee. 2009. Opinion mining and sentiment analysis. *Comput. Linguist* 35, 2 (2009), 311–312.

[37] Souneil Park, Minsam Ko, Jungwoo Kim, Ying Liu, and Junehwa Song. 2011. The politics of comments: predicting political orientation of news stories with commenters' sentiment patterns. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. 113–122.

[38] Juan Ramos. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, Vol. 242. 133–142.

[39] Yanghui Rao, Jingsheng Lei, Liu Wenyin, Qing Li, and Mingliang Chen. 2014. Building emotional dictionary for sentiment analysis of online news. *World Wide Web* 17, 4 (2014), 723–742. Publisher: Springer.

[40] Zhancheng Ren, Qiang Shen, Xiaolei Diao, and Hao Xua. 2021. A sentiment-aware deep learning approach for personality detection from text. 58, 3 (2021), 102532.

[41] Filipe Nunes Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Fabrício Benevenuto, and Marcos André Gonçalves. 2016. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. arXiv:1512.01818 [cs.CL]

[42] Joni Salminen, Hind Almerekhi, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J. Jansen. 2019. Online Hate Ratings Vary by Extremes: A Statistical Analysis. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval* (New York, NY, USA, 2019) *(CHIIR '19)*. ACM, 213–217. https://doi.org/10.1145/3295750.3298954 event-place: Glasgow, Scotland UK.

[43] Joni Salminen, Ahmed Mohamed Kamel, Soon-Gyo Jung, and Bernard Jansen. 2021. The Problem of Majority Voting in Crowdsourcing with Binary Classes. In *Proceedings of 19th European Conference on Computer-Supported Cooperative Work* (Zurich, Switzerland, 2021). European Society for

Socially Embedded Technologies (EUSSET). https://doi.org/10.18420/ecscw2021_n12 Accepted: 2021-05-18T10:05:05Z Publisher: European Society for Socially Embedded Technologies (EUSSET).

[44] David A. Siegel. 2010. The Mystique of Numbers: Belief in Quantitative Approaches to Segmentation and Persona Development. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (2010) *(CHI EA '10)*. ACM, New York, NY, USA, 4721–4732. https://doi.org/10.1145/1753846.1754221 event-place: Atlanta, Georgia, USA.

[45] Robyn Speer and Joanna Lowry-Duda. 2017. Conceptnet at semeval-2017 task 2: Extending word embeddings with multilingual relational knowledge. *arXiv preprint arXiv:1704.03560* (2017).

[46] Maite Taboada. 2016. Sentiment analysis: an overview from linguistics. *Annual Review of Linguistics* (2016). Publisher: Annual Reviews.

[47] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics* 37, 2 (2011), 267–307. Publisher: MIT Press.

[48] Wei Tan, Yushun Fan, Ahmed Ghoneim, M. Anwar Hossain, and Schahram Dustdar. 2016. From the service-oriented architecture to the Web API economy. *IEEE Internet Computing* 20, 4 (2016), 64–68.

[49] Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[50] Mike Thelwall. 2018. Gender bias in sentiment analysis. *Online Information Review* 42, 1 (2018), 45–57.

[51] Peter D. Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *arXiv preprint cs/0212032* (2002).

[52] Svitlana Volkova and David Yarowsky. 2014. Improving gender prediction of social media users via weighted annotator rationales. In *NIPS 2014 Workshop on Personalization*.

[53] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.

[54] Jianwei Zhang, Yukiko Kawai, Shinsuke Nakajima, Yoshifumi Matsumoto, and Katsumi Tanaka. 2011. Sentiment Bias Detection in Support of News Credibility Judgment. In *2011 44th Hawaii International Conference on System Sciences* (2011). 1–10. https://doi.org/10.1109/HICSS.2011.369 ISSN: 1530-1605.

[55] Wenbin Zhang and Steven Skiena. 2010. Trading Strategies to Exploit Blog and News Sentiment. In *Proceedings of the International AAAI Conference on Web and Social Media*.

[56] Huiliang Zhao, Zhenghon Liu, Xuemei Yao, and Qin Yang. 2021. A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach. 58, 5 (2021), 102656.

[57] Ling Zhao, Ying Liu, Mingyao Zhang, and Tingting Guob amd Lijiao Chen. 2021. Modeling label-wise syntax for fine-grained sentiment analysis of reviews via memory-based neural model. 58, 5 (2021), 102641.

Received April 2021; revised July 2021; accepted September 2021