

# Cloud-Based Sentiment Analysis for Interactive Agents

Merel Keijsers

Christoph Bartneck

merel.keijsers@pg.canterbury.ac.nz

christoph.bartneck@canterbury.ac.nz

HIT Lab NZ, University of Canterbury

Christchurch, New Zealand

Hussain Syed Kazmi

hussain.kazmi@enervalis.com

KU Leuven University

Leuven, Belgium

## ABSTRACT

Emotions play an important role in human-agent interaction. To realise natural interaction it is essential for an agent to be able to analyse the sentiment in users' utterances. Modern agents use a distributed service model in which their functions can be located on any number of computers including cloud-based servers. Outsourcing the speech recognition and sentiment analysis to a cloud service enables even simple agents to adapt their behaviour to the emotional state of their users. In this study we test whether sentiment analysis tools can accurately gauge sentiment in human-chatbot interaction. To that effect, we compare the quality of sentiment analysis obtained from three major suppliers of cloud-based sentiment analysis services (Microsoft, Amazon and Google). In addition, we compare their results with the leading lexicon-based software, as well as with human ratings. The results show that although the sentiment analysis tools agree moderately with each other, they do not correlate well with human ratings. While the cloud-based services would be an extremely useful tool for human-agent interaction, their current quality does not justify their usage in human-agent conversations.

## CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; • **General and reference** → *Validation*; *Evaluation*; • **Human-centered computing** → *Web-based interaction*; *Natural language interfaces*; • **Software and its engineering** → *Software evolution*; • **Computer systems organization** → **Embedded systems**; *Redundancy*; *Robotics*; • **Networks** → *Network reliability*.

## KEYWORDS

sentiment analysis; emotion detection; affect recognition; cloud; human agent interaction

## ACM Reference Format:

Merel Keijsers, Christoph Bartneck, and Hussain Syed Kazmi. 2019. Cloud-Based Sentiment Analysis for Interactive Agents. In *Proceedings of the 7th International Conference on Human-Agent Interaction (HAI '19)*, October 6–10,

2019, Kyoto, Japan. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3349537.3351883>

## 1 INTRODUCTION

Affect is an important communication channel in human-agent interaction. Social agents that are able to express sentiment can evoke a range of affective reactions in their human counterparts [see for example 4, 20]. However, for an agent to be able to express the right sentiment at the right time, its affective model needs to perceive as well as interpret the sentiment expressed by the human accurately in the sense of the affective value of events [2].

In the last few years several major software companies, such as Google, IBM, Microsoft or Amazon, have released technologies that not only recognise human speech, but also recognise the sentiment in those utterances. Today, many of these technologies are available as cloud-based services which give agents instant and effortless access to sentiment analyses. This saves not only time and money, but also grants access to algorithms that have been trained on a much larger lexicon than researchers could possibly assemble themselves.

These sentiment analysis services are already utilised for a wide range of tasks, including the automatic detection of sentiment in Twitter feeds or customer feedback postings, which helps companies to determine the general attitude of users who discuss their products and services. In addition, they may prove to be useful for human-AI interaction. Chatbots are a quickly developing field and they have moved past being merely a form of online entertainment. Several technology companies, including Facebook<sup>1</sup> and IBM<sup>2</sup>, offer platforms where other companies can create their own chatbots which can be used, for example, to provide customer support. Other chatbots, such as Siri and Alexa, allow users to verbally ask general questions but also to shop in the connected online shops. In addition, chatbots have been embodied into robots to create personal assistants, like the Jibo and Nabaztag robots.

However, in order to make the interaction with a chatbot truly natural the chatbot needs to be able to recognise the emotions of the speaker, manage its own emotional status and express this status through its voice[2]. This entire system of perceiving, processing, and expressing affect depends on the reliability and validity of the underlying sentiment analysis. A malfunctioning sentiment analysis could potentially do more harm than having no sentiment analysis at all, since it creates the illusion of a conversation partner who is capable of processing and expressing affect, but does not care to respond appropriately [10, 34].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

HAI '19, October 6–10, 2019, Kyoto, Japan

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6922-0/19/10...\$15.00

<https://doi.org/10.1145/3349537.3351883>

<sup>1</sup><https://messenger.fb.com>

<sup>2</sup><https://developer.ibm.com/code/topics/chatbot/>

In this paper we will address the first part of the agents' affect processing: recognising the speakers' affect. While this can be done by the analysis of certain parameters of the speakers' voice [35], we will focus on the analysis of affect based on content analysis only.

A number of major software companies offer services for natural language processing (NLP). These range from speech recognition and sentiment analysis to dialogue management and text-to-speech functions. The question that remains is if these services can be applied to real-life conversational data and how accurately they interpret the sentiment of the users in a conversation. In this paper we investigate the performance of the sentiment analysis of human-agent conversations by three major cloud service companies:

- Google (through its Cloud Natural Language API<sup>3</sup>)
- Amazon (through its AWS Comprehend API<sup>4</sup>)
- Microsoft (through its Text Analytics API<sup>5</sup>)

We benchmark the sentiment detected through these cloud services against two well established methods. First, we compare the results of these cloud-based services on sentiment analysis with arguably the most established stand-alone sentiment analysis software, Linguistic Inquiry and Word Count or LIWC'15 [27, 28]. LIWC'15 makes use of a completely different algorithm for analysing sentiment in a body of text and therefore provides an interesting benchmark for comparison with the performance of the cloud-based services. LIWC was consistently among the best performers in a detailed comparison of sentiment analysis methods [31]. It has been successfully used to detect suicide tendencies in Twitter data [7], and assess personality traits [3, 13] as well as assert general well-being through Facebook data [36], but has, to our knowledge, never been used on human-machine conversation data. We will use LIWC's latest (2015) version.

Second, we compare the results of both the cloud-based services and the lexicon-based software against human interpretation of the sentiment for the same utterances, which we regard as the golden standard. Interestingly, this interpretation can also be performed through a cloud-based service using an API, in this case Amazon Mechanical Turk (MTurk)<sup>6</sup> although other services, such as Crowd-Flower, could also be used. It has been shown that results obtained through MTurk for analysing affect are equivalent to results obtained through lab experiments [1]. We can therefore be confident that the results received through the MTurk API represent human ratings.

## 1.1 Previous work

**1.1.1 Sentiment analysis.** Sentiment analysis refers to a wide array of techniques used to analyse the affect expressed in speech or text. In textual documents, sentiment analysis can be done at multiple levels of detail: document-level, sentence-level and aspect-level [25]. Document-level and sentence-level analysis identify the sentiment of an entire document and a single sentence, respectively. In doing so, they share many common features as sentence-level sentiment analysis can form the basis of document-level analysis. Aspect-level

analysis, on the other hand, is more abstract and tries to identify the affect expressed for different entities (e.g. topics, or specific aspects of the topics) in the text.

Sentiment analysis has been applied in recent years to many different domains. These include the analysis of social network data [5, 23, 40, 41] as well as categorising product reviews and deriving insights about which product features engage users positively and induce them to make purchases [12, 19, 24]. These insights can be useful for sales agents, but similar functionality can also be leveraged by personal assistant agents.

Sentiment analysis tools usually make use of either lexical-based methods or machine learning methods [31]. Lexical-based methods involve calculating the affective charge or orientation for a document from the semantic orientation (e.g. positive, negative or neutral) of words or phrases in the document [38]. These methods perform extremely well when the target documents belong to a similar corpus as the training set. However, a major shortcoming of these methods is their performance on documents which contain many words outside the labelled lexicon. As the method has memorised and not learned which sentiments correspond to which utterances, labelled affect can be incorrect in the case where the training and test data sets differ substantially. With the continuous spatio-temporal evolution of language, it is often necessary to update both the lexicon and the sentiments it identifies. The LIWC'15 sentiment analysis software used in this paper is an example of a classical lexical sentiment analysis method.

Machine learning or statistical methods for sentiment analysis mostly make use of supervised learning algorithms where a classifier is trained on a labelled text corpus [26] to automatically detect patterns in the data. The performance of these models is influenced by both the quality of the training corpus and the capabilities of the learning algorithm. Through observing a large amount of training data, the classifier is expected to be able to generalise to new, unseen examples. This marks an improvement over classical lexical analysis methods as it can be considered closer to the way humans perform sentiment analysis (i.e. through action and interaction).

Recently, deep learning-based methods, a specific form of machine learning, have achieved state-of-the-art performance in natural language processing [17]. Researchers have also applied deep learning models to sentiment analysis, using both traditional text documents and visual data as input [16, 37]. The idea is to learn a general representation underlying the training data to enable the analyser to better generalise to previously unseen data. The ability of the learning method to discover this representation presupposes the existence of a coherent structure underpinning human communication. More concretely, it allows the machine learning system to identify close analogues and relationships between words and phrases, thereby letting it extrapolate to previously unseen cases (for example, man is to king as woman is to queen) [6]. Learning these representations (or word embeddings in natural language processing jargon) has brought about state-of-the-art results in most language tasks including sentiment analysis [8, 39].

Some examples of this recent success can be seen on a number of different sentiment analysis tasks. These include accuracy exceeding 90% on binary sentiment classification (i.e. classifying sentiments to either positive or negative) on a host of different data sets, including the IMDB data set (which consists of 50,000 movie

<sup>3</sup><https://cloud.google.com/natural-language/>

<sup>4</sup><https://aws.amazon.com/comprehend/>

<sup>5</sup><https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/>

<sup>6</sup><https://www.mturk.com/>

reviews), the Stanford Sentiment Treebank data set (which consists of over 200,000 phrases extracted from movie reviews) and the Yelp data set (which consists of over 500,000 Yelp reviews) [18, 22, 30]. There has even been substantial progress recently in learning from one data set and extending it to other domains. This has enabled the accuracy of binary sentiment classification tasks to exceed 80% in test conditions which approximate real world challenges (for instance, a sentiment analyser trained on book reviews being applied to movie reviews etc.) [32]. However, in more nuanced settings where sentiment classification is not binary, as most human interaction is, the performance of these algorithms still degrades considerably. For instance, classifying sentiments on a five-way scale reduces the accuracy to almost 55% on the same Stanford Sentiment Treebank data set [29] and 70% on the Yelp data set [22]. Furthermore, much existing work focuses exclusively on product reviews where it is easy to pinpoint the sentiment expressed (for instance as a rating on a numeric scale). This is not always possible in human-machine interaction where many different types of affect can manifest in a single conversation.

More recently, deep learning-based algorithms have also been applied to the cloud sentiment analysis tools being considered in this paper. All three services we compare (Google, Amazon and Microsoft) reportedly make use of these methods. As such proprietary frameworks reflect the general state of the art, it is plausible that they continue to suffer from the same limitations highlighted above. Furthermore, while being completely data-driven makes these algorithms generalisable, it also leaves them susceptible to systematic bias. For instance, it has been shown that word embeddings trained directly on Google News learn disturbingly stereotypical analogies (e.g. man is to computer programmer as woman is to homemaker) [6]. It is arguably undesirable for agents to perpetuate societal stereotypes in their interactions with humans.

[31] already identified the need to evaluate different lexical and machine learning-based sentiment analysis tools, and have compared a number of different methods for sentiment analysis, covering both lexical and machine learning algorithms. Data sets used for the comparison included product reviews as well as comments on social media and news items. However, none of these were based on the conversational aspect central to human-agent interaction. Thus, the usefulness of sentiment analysis tools in human-agent interaction remains uncertain. In this paper we therefore focus on transcripts of human-bot interactions.

**1.1.2 Communication modalities.** Affect is a complex, multi-modal phenomenon manifested across multiple distinct but interacting modalities [21]. In text-based conversations, which lack physical cues of expression, affect can be conveyed by humans in varying levels of nuance. For example, at one extreme is the use of sarcasm and damnation, like using faint or exaggerated praise; at the other end are specific language cues such as the use of all caps or emoticons. An agent's lack of understanding of these emotional cues can shatter the illusion of reality in human-agent interaction and damage the quality of the conversation. For example, previous research has indicated that people consider a non-empathetic robot more dependable, more credible, and closer to themselves than a robot that is inaccurate in its display of affective responses [10]. Utilising a sentiment analysis tool that fails to correctly identify affective

content may lead to an agent that gives inappropriate replies and is perceived as less trustworthy and relatable than when it had not used any form of sentiment interpretation.

This is an important problem to address as the availability of off-the-shelf API packages enables practitioners to perform sentiment analysis without being aware of the potential limitations. As we have already stressed, a tool's performance on one particular type of language data does not guarantee that it will work as expected on other language data. The accuracy of the sentiment analysis of Twitter data, for example, can be very different compared to the accuracy of processing comments to newspaper articles [31]. Since both the algorithms and the training corpus behind the cloud-based systems remain undisclosed, practitioners are encouraged to treat the online sentiment analysis tool as a black box. This can lead to unintended consequences such as an overconfidence in potentially dubious and uncertain results.

## 1.2 Research Questions

The primary research question we intend to answer is to what degree different sentiment analysis tools, both cloud-based and lexical, agree with human ratings as well as with each other, when used to analyse human-machine conversation data. More precisely, the research questions are:

- (1) Do the cloud-based sentiment analysis tools agree with each other?
- (2) Do the cloud-based sentiment analysis tools agree with the lexicon-based software?
- (3) Do the cloud-based sentiment analysis tools agree with human judgements?
- (4) Does the lexicon-based software output agree with human judgements?
- (5) To what extent do human judgements of the same conversation agree with one another?

## 2 METHOD

In this paper we extend the methodology introduced in [31] by focusing specifically on conversations between humans and machines. Another key difference is our focus on document-level sentiment analysis as opposed to their sentence-level analysis. Furthermore, while a majority of software methods they employed made use of lexical analysis methods, we consider machine learning-based methods provided by large cloud service providers.

### 2.1 Material

**2.1.1 Conversation data.** We are unaware of any substantial openly available speech corpus from humans interacting with agents so we use the possibly next best speech data between humans and machines. In this paper 285 anonymous conversations between Cleverbot, an online chatbot, and a user are analysed. Cleverbot is based on the award-winning Jabberwacky engine. Cleverbot does not make use of any state-based machines or scripts to carry out its conversation. Rather, it repeats things previous human users have said to it in similar context. This can produce either extremely realistic or bizarre conversations. We obtained a set of anonymous conversations directly from the Cleverbot owner who collected them between 12pm and 12.30pm, and a second between 12am and

12.30am (both UK time) on 17 April 2018. Only conversations which contained 20 turns or more and were held in English were included in the analysis.

The input data set was organised as a dialogue between a human user and Cleverbot. Before being put through the sentiment analysis tools, the tags of “User” and “Cleverbot” were removed to avoid any bias. For the human raters, they were replaced with “User1” and “User2”. Again, this was done to avoid bias. Therefore the sentiment analysis of the conversations should be indifferent to whether it was machine generated and/or included human participation.

**2.1.2 Cloud Service Tools.** Running sentiment analysis with the cloud-based platforms allows us to identify the affect expressed in a conversation. All the APIs can be used for both sentence-level and document-level analysis. We used document-level since it analyses the complete conversation and not just a single utterance. As has already been noted in literature, different platforms return affect on different scales [31].

Amazon, through its AWS Comprehend API, returns four different types of affect: Positive, Negative, Neutral, and Mixed. Each type is scored in the range of 0 and 1, and the sum is always equal to 1.

Google, through its Natural Language Processing API, returns two different numbers: Score and Magnitude. It thereby relates to the popular circumplex model of affect [33]. The score ranges between -1 and 1 where higher numbers reflect more positive affect. The magnitude can range between 0 and  $\infty$ , and refers to the strength of affect expressed in the text. This two-dimensional output means that a score around 0 and a low magnitude reflects a mostly neutral document, while a score of 0 combined with a high magnitude reflects a mixed document.

Microsoft, through its Text Analytics API, returns only a single sentiment score which is scaled between 0 and 1. Higher scores correspond to more positive sentiments in the document.

**2.1.3 Lexicon-based analysis tool.** Unlike cloud-based sentiment analysis, the Linguistic Inquiry and Word Count (LIWC'15) software [27] works with a dictionary which is used to sort and sift through words and phrases in a text. It returns scores on a wide range of variables, from percentage of words over 6 letters and use of pronouns, to informal language markers and psychological constructs (e.g. “cognition” and “affect”). In the current study the following scales are used: “affect” (the mention of any affective processes in the text), which is divided in the subscales Positive affect and Negative affect, the latter of which in turn has the subscales Anxiety, Anger, and Sadness; and, in addition to these, the scale “swear” which reflects use of swear words. Each of these scales range from zero to  $\infty$ . Although the subscales Positive and Negative affect add up to the Affect score, the sum of the Anxiety, Anger, and Sadness subscales does not necessarily add up to the Negative affect; i.e. the Negative affect scale is more than just the sum of its three subscales.

**2.1.4 Human ratings.** Amazon Mechanical Turk (MTurk) was used to collect human ratings of each conversation. Each conversation was set out to be rated by four different raters on each of the four Amazon scales: Positive, Negative, Neutral, and Mixed, on a five-point Likert scale (e.g. for the Positive scale raters could opt between

“Not Positive”, “Slightly Positive”, “Positive”, “Very Positive”, and “Extremely Positive”). These four scales were chosen as they translate relatively well to the scales of the different sentiment analysis tools.

Overall, 63 MTurk workers rated on average 16.73 conversations each ( $SD = 35.58$ ,  $\min = 1$ ,  $\max = 174$ ). Some data loss occurred during the conversion of the conversations into a format that could be read by MTurk, resulting in a total of 279 conversations being rated. In addition, ratings that took less than 25 seconds were removed from the data set as it seemed unlikely that the rater had fully read and rated the conversation in such a short time. In the end, 230 conversations were rated four times, 48 were rated three times, and one was rated only twice.

No demographics were collected on the raters. To ensure the quality of the data, participation was restricted to workers from English speaking countries (thus warranting that they were proficient enough to understand the texts they were asked to rate) and only workers with a ‘Master worker’ status were allowed to contribute. Master workers are identified by an MTurk algorithm as workers who provide high quality work across a wide variety of tasks. Moreover, if a Master worker’s performance drops below the threshold, their title as Master is revoked.

### 3 RESULTS

The comparison of the different software analysis tools could take advantage of the full 285 conversations, while any comparisons to the human ratings were constrained to the 279 conversations available. See Table 1 for the descriptives of the words per sentence, word count, and turn count.

After an exploratory sentiment analysis by the cloud service tools on sentence level gave noisy and random results, with no discernible patterns arising over the course of the conversations, the granularity of the sentiment analysis was set to the conversation level.

#### 3.1 Agreement between the cloud service tools

To make the interpretation easier, the cloud service data was transformed so all sentiments were reported on a range from 0 to 1 with 0 being negative, 1 being positive, and .5 being neutral. Microsoft’s Text Analysis tool already reports scores in this range, so no transformation was required. Google’s Natural Language Processing tool reports on a -1 to 1 scale, which meant it was a matter of normalising the score without taking the magnitude into account. Amazon’s AWS Comprehend tool reports four scores; of these, only the positive and negative sentiment were considered. The negative was subtracted from the positive and the result was added to .5 to derive a single score in the range of 0 and 1. A visual

**Table 1: Descriptives of the conversations**

	Mean ( <i>SD</i> )	Median	67% interval
Words per sentence	4.36 (2.65)	4.00	[2.96; 4.92]
Turn count	94.98 (44.04)	88.00	[64.00; 115.54]
Word count	573.70 (262.32)	514.50	[381.46; 748.54]

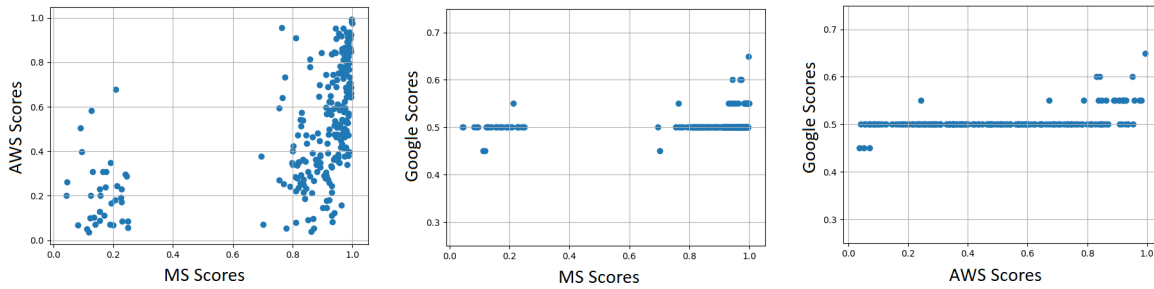


Figure 1: Scatter plots of the different sentiment scores. From top to bottom: Amazon (AWS) scores against Microsoft scores; Google scores against Microsoft scores; Google scores against Amazon (AWS) scores.

Table 2: Correlation coefficients (Kendall's tau) between the LIWC'15 lexical analysis and the cloud service tools

		LIWC'15 scales						
		Affect	Positive affect	Negative affect	Anxiety	Anger	Sadness	Swearing
Amazon	Positive	.284	.410	-.160	<i>n.s.</i>	-.158	<i>n.s.</i>	-.164
	Negative	<i>n.s.</i>	-.265	.416	<i>n.s.</i>	.420	.107	.294
	Mixed	-.279	-.171	-.212	<i>n.s.</i>	-.238	-.096	-.159
	Neutral	.226	.140	.228	<i>n.s.</i>	.228	.110	.156
Google	Feeling	.264	.349	-.186	<i>n.s.</i>	-.213	<i>n.s.</i>	-.127
	Intensity	.104	.106	<i>n.s.</i>	.103	.105	.121	.089
Microsoft	Sentiment score	.106	.305	-.342	<i>n.s.</i>	-.308	-.170	-.271

inspection of the data as shown in Figure 2 indicates that the three APIs show very different results even when normalised to the same scale. Most notably, the scores from the Google API are clustered around 0.5 which, in conjunction with mostly high magnitudes, means it has detected very mixed affect. This is in stark contrast with the scores of Amazon (AWS) and Microsoft. Microsoft seems to be applying a possibly non-linear function to the output of its sentiment analysis to force scores either close to 0 or 1. As a result, it primarily classifies the conversations as positive, with a sizeable minority in the negative sentiment. Normalised sentiment scores from Amazon appear to be spread fairly uniformly across the whole range between 0 and 1.

Kendall's  $\tau$  is used as a measure for correlation due to the non-normal distribution of Microsoft's results. For Amazon and Microsoft scores  $\tau$  is 0.49 ( $p < .001$ ), which can be considered a large effect [9, 15]. Google's results correlate less well with both Microsoft ( $\tau = .26$ ,  $p < .001$ , medium effect) and Amazon ( $\tau = .33$ ,  $p < .001$ , medium effect) as it squashes all results close to 0.5. Figure 1 shows the scatter plots of the three cloud services.

### 3.2 Agreement between cloud service tools and LIWC'15

As can be seen in Table 2, the LIWC'15 and most cloud service analysis results correlated reasonably well.

*Amazon.* Although most correlations are both present and in the direction that could be expected, (for example, positive affect as identified by Amazon's Comprehend analysis tool correlating positively with Positive affect from LIWC'15) some correlations, or lack of correlations, are quite surprising. One of those is the

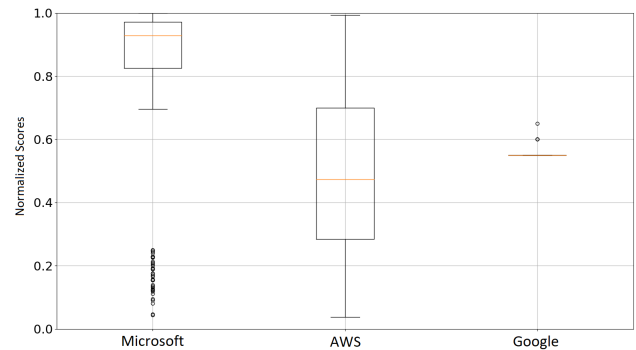


Figure 2: Distribution of sentiments (with 1 being very positive, 0 being very negative, and .5 being neutral) according to the different APIs

positive correlation between Amazon's Neutral scale and nearly all of LIWC'15's affect scales, suggesting that as Amazon identified a conversation as more Neutral in tone, LIWC'15 rated it as containing more affect. Conversely, a negative correlation between Mixed affect by Amazon and LIWC'15's rating of Affect indicates that as Amazon identified a conversation as being more mixed in its sentiment, LIWC'15 identified it as containing less affect. Moreover, the lack of significant correlation between the Positive and Negative Amazon scales on one hand and LIWC'15's Anxiety subscale on the

other hand, suggests that Amazon may do well in identifying general positive and negative affect, but does not discern well between the different kinds of negative affect.

**Google.** The analysis outcomes of Google’s Natural Language Processing tool correlated well with the LIWC’15 analysis results. Interestingly, Anxiety as identified by LIWC’15 did not correlate with Google’s Feeling scale.

**Microsoft.** The Sentiment score of Microsoft’s Text Analysis tool correlated well with the LIWC’15 results. Like the sentiment analysis scales of Amazon and the Feeling scale of Google, Microsoft’s Sentiment score did not correlate with Anxiety. This pattern of non-significant correlations suggests that either the cloud services do not identify anxious sentiment well, or that the LIWC’15 does not detect actual anxiety.

### 3.3 Agreement between tools and human judgement

#### 3.3.1 Cloud service tools.

**Amazon.** Surprisingly, there was a marginally significant negative correlation between the Positive sentiment scale of Amazon’s Comprehend analysis tool and human judgement of how positive a conversation was ( $\tau = -.078$ ,  $p = .058$ , small to medium effect). In addition, the human Negative sentiment ratings correlated negatively with Amazon’s Mixed sentiment ( $\tau = -.093$ ,  $p = .025$ , medium effect) and marginally with Neutral sentiment ( $\tau = .075$ ,  $p = .071$ , small to medium effect). This suggests that the Amazon sentiment analysis is not capable of correctly identifying any sentiments in human-machine conversations. In the best case, its results are unrelated to human judgement and in the worst case they are diametrically opposed.

**Google.** The outcomes of Google’s Natural Language Processing service did not correlate very well with human judgement either. The sole correlation that was marginally significant was a curiously positive correlation between Google’s Feeling scale and the human judgement of Neutral sentiment ( $\tau = .095$ ,  $p = .063$ ). These scores suggest that the Google cloud service sentiment analysis tool is not capable of correctly identifying any sentiments in human-machine conversations.

**Microsoft.** The single sentiment score that was returned by Microsoft’s Text Analytics tool correlated positively with the Neutral judgement by the human raters ( $\tau = .094$ ,  $p = .041$ , medium effect), suggesting that it is not capable of correctly identifying the accurate sentiments in human-machine conversations.

**3.3.2 LIWC’15.** As can be seen in the bottom half of Table 3, the different sentiment scales of the LIWC’15 did not correlate with the human judgements on many points. The Affect scale of LIWC’15 correlated positively ( $\tau = .089$ ,  $p = .036$ , medium effect) with the Mixed scale of the human judgements. There were negative correlations between the human judgement of Neutrality on one hand and the LIWC’15 scale Negative Affect ( $\tau = -.127$ ,  $p = .003$ , medium effect) and its subscales Anxiety ( $\tau = -.105$ ,  $p = .025$ , medium effect) and Anger ( $\tau = -.100$ ,  $p = .019$ , medium effect). Moreover, a positive correlation was found between the LIWC’15 Anxiety subscale and

**Table 3: Correlation coefficients (Kendall’s  $\tau$ ) between human judgement and the scales of the cloud service tools and the LIWC’15 lexical analysis.**

		Human judgements			
		Positive	Negative	Mixed	Neutral
Amazon (AWS)	Positive	-.078 <sup>†</sup>	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>
	Negative	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>
	Mixed	<i>n.s.</i>	-.093	<i>n.s.</i>	<i>n.s.</i>
	Neutral	<i>n.s.</i>	.075 <sup>†</sup>	<i>n.s.</i>	<i>n.s.</i>
Google	Feeling	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>	.095 <sup>†</sup>
	Intensity	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>
Microsoft	Sentiment	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>	.094
LIWC’15	Affect	<i>n.s.</i>	<i>n.s.</i>	.089	<i>n.s.</i>
	Positive	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>
	Negative	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>	-.127
	Anxiety	<i>n.s.</i>	.103	<i>n.s.</i>	-.105
	Anger	<i>n.s.</i>	<i>n.s.</i>	.071 <sup>†</sup>	-.100

<sup>†</sup> indicates a marginally significant ( $.05 < p < .10$ ) correlation

the human judgement of Negative affect ( $\tau = .103$ ,  $p = .027$ , medium effect), as well as a marginally significant correlation between the Anger subscale of the LIWC’15 and the human judgement of Mixed sentiments ( $\tau = .071$ ,  $p = .097$ , small effect). Neither the Sadness subscale nor Swearing scale of the LIWC’15 showed any significant correlation with human judgements, so were not included in Table 3.

The combination of the (negative) correlations between the human judgement of a neutral tone in the conversations and LIWC’15’s detection of Negative affect, and the (positive) correlation between the Affect and Anger scales and the human judgement of anger, suggest that the tool was able to pick up on some sentiment but couldn’t correctly identify any of the affect save for Anxiety.

### 3.4 Agreement amongst human judgements

As stated in section 2.1.4, each conversation was scored by four (or sometimes three) different human raters; these raters were not identical for all conversations. As some raters completed only a handful of ratings, and others finished many, inter rater agreement cannot be computed. In order to still give some indication of agreement, the standard deviation between ratings on each scale (i.e. Positive, Negative, Neutral and Mixed) within each unique conversation was calculated. This standard deviation statistic indicates how close the ratings on each particular scale for each individual conversation were to one another. For example, if the four raters of a single conversation had rated it 2, 3, 4, and 3 out of 5 on Negativity respectively, the standard deviation of the Negative scale would be .82. The higher the standard deviation, the less the raters agreed with one another.

## 4 DISCUSSION

The current paper tested the usability of different sentiment analysis tools for conversational data, by measuring to what extent the tools agreed with both human raters and each other on the sentiment in human-bot conversations. The cloud-based sentiment analysis tools

agreed reasonably well with one another and the lexical analysis software. However, the sentiment analysis tools did not correlate with human ratings, suggesting that although the tools picked up on some cues in the text, they either missed other important cues, or misinterpreted the cues they picked up, or both.

The human raters largely agreed among each other on the positive and negative scales, which suggests that the stimuli used in our study are sufficiently unambiguous. The different software tools, however, were unable to detect the sentiment patterns in the conversations. While it is hardly surprising that humans are better at detecting sentiment than software, the size of the disparity casts considerable doubt on the reliability and validity of the sentiment analysis tools currently offered for use in human-agent interaction.

For example, the conversation below was rated as *mostly mixed* by Amazon, *neutral* and *not very intense* by Google, and *very positive* by both Microsoft and LIWC'15.

CLEVERBOT: What do you want to do?

USER: I want to slit your face open with a hot knife.

CLEVERBOT: So do I, dingbat!

USER: What ever.

CLEVERBOT: Do you live in America?

USER: Nope.

CLEVERBOT: Good. I would've asked to move there.

USER: I live in africa and i have ebloa [sic].

CLEVERBOT: Is that where your pineapple is?

USER: Yes.

CLEVERBOT: I take it you like it there.

USER: Yuppo.

CLEVERBOT: Why not?

USER: Ee your [sic] too irrelevant.

CLEVERBOT: No, it's not honey.

USER: Byew bitcgh [sic] v.

CLEVERBOT: What are you talking about?

USER: Your arse.

In contrast, human raters judged this conversation as very negative, slightly mixed, not positive, and not neutral.

Since the cloud service tools are not open about how their algorithms were developed, it is hard to determine what caused the discrepancy between human and tool ratings. One possibility would be that the service tools are mainly trained on a different type of data (like tweets and reviews) and as a result cannot deal well with the more volatile nature of human-agent dialogue. It is also possible that the internet slang, swearing and sarcasm in the conversations threw the tools off balance. Of course, these explanations are not mutually exclusive. Supporting such explanations, Ribeiro et al. [31] ranked various sentiment analysis tools yet still concluded

that their benchmarks depended considerably on the data set used. Despite rapid progress owing to deep learning, domain adaptation across different data sets remains the panacea for natural language processing algorithms. In short, while sentiment analysis may be tremendously useful on other types of data (e.g. tweets, reviews, Facebook status updates), we would like to caution against the naive use of these services on data types with which they may not be familiar with, such as human-agent interaction and other types of human-machine interaction.

## 4.1 Limitations

The conversations between users and Cleverbot are often strange and jump from one topic to the next. While this makes for quite particular data, we argue that the current analysis is relevant for a broader scope of human-agent interaction as well. As technology progresses and AI becomes a more central part of everyday life, informal conversations between humans and agents will become more ubiquitous - already, people use their Alexa or Siri for a wide range of conversation topics which stretches far beyond mere inquiries about the weather or news headlines [11, 14]. While a recreational bot like Cleverbot may be less sophisticated than an Alexa, we expect there to be at least some overlap in the conversations people will attempt to have with either agent. Thus, the inability of the sentiment analysis tools to correctly identify affect in quirky, unpredictable conversations is an issue not just restricted to Cleverbot data.

Another point to consider is that our analysis used utterances from both users and Cleverbot, who at times could be inconsistent. It could be argued that a more favourable image of the sentiment analysis tools would have emerged if solely the utterances of the user had been taken into account. However, as also shown in the conversation snippet above, humans often change the topic of conversation at will. The sentiment analysis tools discussed in this paper are designed to analyse discussion in online forums and social media commentary. They should therefore be designed to take contributions from multiple, occasionally inconsistent, conversation partners into account. We therefore do not believe that the current study disadvantaged the sentiment analysis tools by feeding them the full conversations.

## 4.2 Future work

When making sense of what the conversation partner is trying to convey, humans use a wide range of cues, like facial expressions, volume of speech, and body language, but also less obvious cues like personality or how someone dresses. Sentiment analysis therefore could take not only the bare text into account but also consider the voice of the speaker, self-corrections and hesitations in the speech, and all other non-verbal cues. Ideally, the cloud-based services would expand their service to also include non-verbal data in order to improve the quality of their analyses.

**Table 4: Descriptives of the standard deviations between ratings of the same conversation, per scale.**

Scale	Median	Mean	67% interval
Positive	.71	.75	[.58; .96]
Negative	.58	.72	[.50; .96]
Neutral	.96	.98	[.82; 1.15]
Mixed	.82	.83	[.58; .96]

## REFERENCES

- [1] Christoph Bartneck, Andreas Duenser, Elena Moltchanova, and Karolina Zawieska. 2015. Comparing the similarity of responses received from experiments conducted in Amazon's Mechanical Turk to experiments conducted with traditional methods. *PLOS One* 10, 4 (2015), e0121595. <https://doi.org/10.1371/journal.pone.0121595>

- [2] Christoph Bartneck and Michael J. Lyons. 2009. Facial Expression Analysis, Modeling and Synthesis: Overcoming the Limitations of Artificial Intelligence with the Art of the Soluble. In *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence*, Jordi Vallverdú and David Casacuberta (Eds.). IGI Global, Chapter 3, 33–53.
- [3] B. Bazelli, A. Hindle, and E. Stroulia. 2013. On the Personality Traits of StackOverflow Users. In *IEEE International Conference on Software Maintenance*. 460–463. <https://doi.org/10.1109/ICSM.2013.72>
- [4] Cindy L Bethel and Robin R Murphy. 2008. Survey of non-facial/non-verbal affective expressions for appearance-constrained robots. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38, 1 (2008), 83–92. <https://doi.org/10.1109/TSMCC.2007.905845>
- [5] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science* 2, 1 (2011), 1–8. <https://doi.org/10.1016/j.jocs.2010.12.007>
- [6] 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.
- [7] Scott R Braithwaite, Christophe Giraud-Carrier, Josh West, Michael D Barnes, and Carl Lee Hanson. 2016. Validating machine learning algorithms for Twitter data against established measures of suicidality. *JMIR mental health* 3, 2 (2016). <https://doi.org/10.2196/mental.4822>
- [8] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).
- [9] Jacob Cohen. 1992. A power primer. *Psychological bulletin* 112, 1 (1992), 155.
- [10] Henriette Cramer, Jorrit Goddijn, Bob Wielinga, and Vanessa Evers. 2010. Effects of (in) accurate empathy and situational valence on attitudes towards robots. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*. IEEE, 141–142.
- [11] Amanda Cercas Curry and Verena Rieser. 2018. # MeToo Alexa: How Conversational Systems Respond to Sexual Harassment. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*. 7–14.
- [12] Xing Fang and Justin Zhan. 2015. Sentiment analysis using product review data. *Journal of Big Data* 2, 1 (2015), 5. <https://doi.org/10.1186/s40537-015-0015-2>
- [13] Golnoosh Farnadi, Susana Zoghbi, Marie-Francine Moens, and Martine De Cock. 2013. Recognising personality traits using Facebook status updates. In *Proceedings of the workshop on computational personality recognition (WCPRI3) at the 7th international AAAI conference on weblogs and social media*. <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6245/6309>
- [14] Leah Fessler. 2017 (accessed 06-March-2019). We tested bots like Siri and Alexa to see who would stand up to sexual harassment. <https://bit.ly/2LGF8w>
- [15] Andrew R Gilpin. 1993. Table for conversion of Kendall's Tau to Spearman's Rho within the context of measures of magnitude of effect for meta-analysis. *Educational and psychological measurement* 53, 1 (1993), 87–92. <https://doi.org/10.1177/0013164493053001007>
- [16] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning*. 513–520. [http://www.icml-2011.org/papers/342\\_icmlpaper.pdf](http://www.icml-2011.org/papers/342_icmlpaper.pdf)
- [17] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*. Vol. 1. MIT press Cambridge.
- [18] Scott Gray, Alec Radford, and Diederik P Kingma. 2017. *GPU kernels for block-sparse weights*. Technical Report. Technical report, OpenAI.
- [19] Emitza Guzman and Walid Maalej. 2014. How do users like this feature? a fine grained sentiment analysis of app reviews. In *Requirements Engineering Conference (RE), 2014 IEEE 22nd International*. IEEE, 153–162. <https://doi.org/10.1109/RE.2014.6912257>
- [20] Louisa Hall. 2017. How We Feel About Robots That Feel. *MIT Technology Review* (2017). <https://www.technologyreview.com/s/609074/how-we-feel-about-robots-that-feel/>
- [21] Eva Hudlicka. 2008. What are we modeling when we model emotion?. In *AAAI spring symposium: emotion, personality, and social behavior*, Vol. 8. AAAI. <http://www.aaai.org/Papers/Symposia/Spring/2008/SS-08-04/SS08-04-010.pdf>
- [22] Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 562–570.
- [23] Young Bin Kim, Jun Gi Kim, Wook Kim, Jae Ho Im, Tae Hyeon Kim, Shin Jin Kang, and Chang Hun Kim. 2016. Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PloS one* 11, 8 (2016), e0161197. <https://doi.org/10.1371/journal.pone.0161197>
- [24] Ting-Peng Liang, Xin Li, Chin-Tsung Yang, and Mengyue Wang. 2015. What in consumer reviews affects the sales of mobile apps: A multifacet sentiment analysis approach. *International Journal of Electronic Commerce* 20, 2 (2015), 236–260. <https://doi.org/10.1080/10864415.2016.1087823>
- [25] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal* 5, 4 (2014), 1093–1113. <https://doi.org/10.1016/j.asej.2014.04.011>
- [26] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 79–86. <https://doi.org/10.3115/1118693.1118704>
- [27] James W Pennebaker, Roger J Booth, Ryan L Boyd, and Martha E Francis. 2015. *Linguistic Inquiry and Word Count: LIWC 2015* [Computer software]. Pennebaker Conglomerates. <http://liwc.wpengine.com>
- [28] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report. The University of Texas at Austin. <http://hdl.handle.net/2152/31333>
- [29] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).
- [30] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444* (2017).
- [31] Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabricio Benevenuto. 2016. SentiBench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science* 5, 1 (2016), 23. <https://doi.org/10.1140/epjds/s13688-016-0085-1>
- [32] Sebastian Ruder and Barbara Plank. 2018. Strong Baselines for Neural Semi-supervised Learning under Domain Shift. *arXiv preprint arXiv:1804.09530* (2018).
- [33] J.A. Russel. 1979. Affective space is bipolar. *Journal of personality and social psychology* 37, 3 (1979), 345–356. <https://doi.org/10.1037/0022-3514.37.3.345>
- [34] Maria Ruz and Pío Tudela. 2011. Emotional conflict in interpersonal interactions. *Neuroimage* 54, 2 (2011), 1685–1691.
- [35] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication* 53, 9 (2011), 1062–1087. <https://doi.org/10.1016/j.specom.2011.01.011>
- [36] Michele Settanni and Davide Marengo. 2015. Sharing feelings online: studying emotional well-being via automated text analysis of Facebook posts. *Frontiers in psychology* 6 (2015), 1045. <https://doi.org/10.3389/fpsyg.2015.01045>
- [37] Aliaksei Severyn and Alessandro Moschitti. 2015. Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 959–962. <https://doi.org/10.1145/2766462.2767830>
- [38] 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. [https://doi.org/10.1162/COLI\\_a\\_00049](https://doi.org/10.1162/COLI_a_00049)
- [39] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1555–1565.
- [40] Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahr, and Cornelius Puschmann. 2014. *Twitter and society*. Vol. 89. Peter Lang Inc.
- [41] Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. 2011. *Combining lexicon-based and learning-based methods for twitter sentiment analysis*. Technical Report 89. HP Laboratories. <https://uic.pure.elsevier.com/en/publications/combining-lexicon-based-and-learning-based-methods-for-twitter-se>