

2020-12-31

Investigating the Lack of Consensus Among Sentiment Analysis Tools

PALOMINO, MARCO

<http://hdl.handle.net/10026.1/16929>

10.1007/978-3-030-66527-2_5

Lecture Notes in Computer Science

Springer Verlag

All content in PEARL is protected by copyright law. Author manuscripts are made available in accordance with publisher policies. Please cite only the published version using the details provided on the item record or document. In the absence of an open licence (e.g. Creative Commons), permissions for further reuse of content should be sought from the publisher or author.

Investigating the Lack of Consensus among Sentiment Analysis Tools

Marco A. Palomino¹[0000–0001–7850–416X], Aditya Padmanabhan Varma²,
Gowriprasad Kuruba Bedala³, and Aidan Connelly¹

¹ University of Plymouth, Plymouth, UK
{marco.palomino,aidan.connelly}@plymouth.ac.uk

² Vellore Institute of Technology, Chennai, India
adityapadmanabhan.2016@vitstudent.ac.in

³ Saveetha Institute of Technical and Medical Sciences, Chennai, India
bedalagowriprasad16@saveetha.com

Abstract. Sentiment analysis, the classification of human emotion expressed in text, has the potential to enhance our ability to analyse the ever growing amount of information published each day on social media. Thus, we compare here seven of the most well-regarded sentiment analysis tools, and conclude that none of them is sufficiently reliable to be used on its own. Combining them and relying on their results only when various tools reach an agreement seems to be a better option. The pros and cons of such an approach are discussed in this paper, while providing recommendations related to the usability of the tools in question. Our work is of particular relevance to small and medium-sized enterprises (SMEs), which constitute a large and integral part of the economy. SMEs seem to be ideal candidates to turn data derived from sentiment analysis into business opportunities.

Keywords: Sentiment analysis · Twitter · social media · scikit-learn · Sentiment140 · SentiStrength · uClassify · VADER · TextBlob · SMEs.

1 Introduction

According to the *European Union*, and other international organisations, such as the World Bank and the United Nations, *small and medium-sized enterprises* (SMEs) are businesses whose personnel falls below 250 employees, and whose annual turnover does not exceed EUR 50 million [13]. In the UK, 5.8 million small businesses were in operation at the start of 2019 [47]. Indeed, small businesses accounted for 99.3% of all private sector businesses—these were 5.82 million businesses with 0 to 49 employees—and SMEs accounted for three fifths of the employment and around half of the turnover in the UK private sector at the start of 2019 [47]. Considering that the total employment in SMEs across the UK is currently 16.6 million, which equates to 60% of all private sector employment, supporting the needs of SMEs has become a critical issue.

The software designed to support the operation of SMEs is meant to help them to run operations, cut costs and replace paper processes [27]. We are particularly interested in the software choices available for a specific application that has been gaining interest and popularity: *sentiment analysis*, the process of computationally categorising opinions [14].

Sentiment analysis—sometimes known as *opinion mining*—aims to systematically identify, extract, quantify, and study opinions about specific topics, and attitudes towards particular entities [4, 14]. Sentiment analysis has a great potential as a technology to enhance the capabilities of customer relationship management and recommendation systems—for example, showing which features customers are particularly happy about, or excluding from recommendations items that have received negative feedback. Sentiment analysis can also be exploited for troll-filtering and spam detection [5]. Intelligence applications able to monitor surges in hostile communications are examples of non-commercial systems employing sentiment analysis [18, 29].

The basic tasks of sentiment analysis are *emotion recognition* [41] and *polarity detection* [21]. While the first task focuses on identifying a variety of emotional states, such as “anger”, “sadness” and “happiness”, the second one is either a binary classification task—whose outputs are ‘positive’ versus ‘negative’, ‘thumbs up’ versus ‘thumbs down’, or ‘like’ versus ‘dislike’—or a ternary classification task—whose outputs are ‘positive’, ‘neutral’ or ‘negative’. Several sentiment analysis tools have been developed lately—both Feldman [14] and Ribeiro *et al.* [39] claim that 7,000 articles on sentiment analysis had been written up by 2016, while dozens of start-ups are developing sentiment analysis solutions.

Despite the interest in the subject, it is still unclear which sentiment analysis tool is more adaptable to different domains, or cheaper and easier to manage. Therefore, the goal of this paper is to help SMEs to evaluate off-the-shelf tools for the purpose of sentiment analysis, and ascertain which tool is better for specific needs that businesses may encounter. Little is known about the relative performance of the various tools available [39]; thus, comparative studies such as this one are needed.

Our initial evaluation suggested that sentiment analysis can be severely biased, depending on which tool is used [8]. Consequently, we launched a larger investigation in 2020, where we have added new sentiment analysis tools to the analysis and used a much larger corpus as a testbed for our experimentation. We can now confirm that considering the consensus among a selection of tools is a better alternative than choosing one and using it in isolation.

The remainder of this paper is organised as follows: Section 2 introduces the corpora for our experiments—we have gathered two different datasets to compare sentiment analysis tools: one in 2017 and one in 2020. Section 3 describes the tools we have compared: *Sentiment140* [16], *SentiStrength* [9], *scikit-learn* [35], *TextBlob* [23], *Treebank* [46], *uClassify* [53] and *VADER* [37]. Section 4 presents the results yielded by the tools we compared and discusses our analysis. Finally, Section 5 offers our conclusions.

2 Experimental Corpus

While large companies can afford time and resources to look into the best sentiment analysis tools for their purposes—for example, IBM acquired *AlchemyAPI* in 2015, before replacing it with the *Watson Natural Language Understanding Service* [11]—most SMEs would find it unreasonable to invest significantly on such an activity. Hence, we decided to launch an investigation of sentiment analysis tools in 2017, as a means to inform SMEs about the features, strengths and drawbacks of popular off-the-shelf sentiment analysis tools.

As a testbed for our experiments, we have chosen *Twitter* [52], the microblogging service that enables people to publish short messages—namely, *tweets*—expressing interests and attitudes they are willing to share [3]. Twitter users employ *hashtags*—words or phrases preceded by a hash sign ‘#’—to categorise tweets topically, so that others can follow conversations on a particular topic. A more detailed description of Twitter and its jargon can be found in [26].

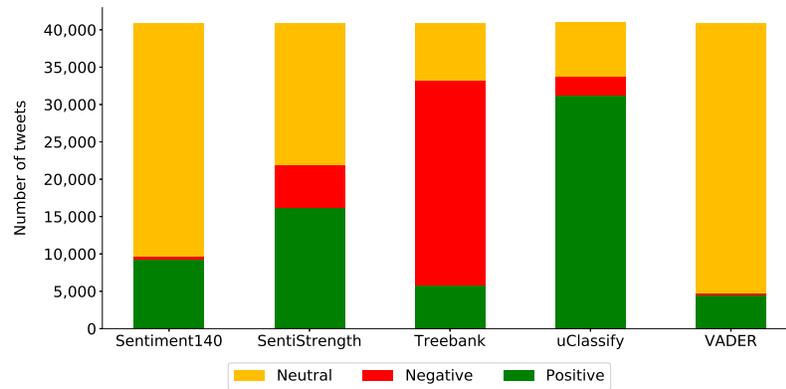
Twitter is a valuable source of opinions and sentiments [31]—for example, manufacturing companies are always interested in how positive or negative the opinions tweeted about their products are. Companies across the world have embraced Twitter as a powerful way to connect with their customers and grow their businesses [7]. Twitter is now indispensable in marketing, sales and customer service. Thus, we have used Twitter for the evaluation of sentiment analysis tools since our first study in 2017.

In 2017, we worked with a corpus consisting of 40,912 tweets collected at the beginning of the year, when people tend to make New Year resolutions. Such resolutions are commonly associated with weight loss and dietary regimes. Hence, this gave us an opportunity to monitor tweets related to nutritional, detox and dietary products. We began the retrieval of the 2017 corpus on 26th January 2017, and we ended it 20 days later—14th February 2017. To guarantee that we gathered a good sample of tweets, a professional in the field provided a list of hashtags and phrases relevant to the subject, which are displayed in Table 1. Such hashtags and phrases captured conversations related to health and disease connected with nutritional and dietary products. Table 1 also displays the number of tweets we collected for each hashtag and phrase. While some of the hashtags seem unintelligible to a layman, they are all sensible in the context of dietary products. For example, *irritable bowel syndrome*—referred to by the hashtag **#IBS**; see row 3 in Table 1—is a condition of the digestive system that is frequently mentioned in dietary conversations. In fact, **#IBS** was the third most popular hashtag in our 2017 corpus.

As explained in [8], the study we carried out in 2017 showed significant differences in the number of tweets classified as positive, negative or neutral, depending on the tool chosen for the classification. Figure 1 displays the polarity of the tweets according to the different tools involved in the study. Such contrasting results led us to undertake further investigation. We are currently studying a greater number of sentiment analysis tools, and we are employing a much larger corpus.

Table 1. Number of tweets per hashtag and phrase in the 2017 corpus.

Hashtag or phrase	Number of tweets
#healthy #food	11,267
#cleaneating	7,853
#IBS	3,974
#foodallergy	3,817
#gluten	3,652
#superfoods	3,556
#lowfodmap	867
#fodmap	829
#natural #diet	546
detox diet nutrition	320
#detoxdiet	224
#diet #research	58
#lowgi	56
#nutraceutical	29
#medicalfood	19
#cleansing #diet	12
#diet #scam	7
food is your medicine	0

**Fig. 1.** Polarity per tool.

Our new 2020 corpus consists of 1,525,050 tweets gathered during 13 continuous hours, starting on Sunday 2nd February 2020 at 20:47:00 (GMT)—hereafter, all times are GMT times. The first tweet was captured at 20:47:03, and the last one at 09:25:08 on Monday 3rd March 2020. Our corpus consists of publicly available tweets referring to the *Super Bowl*, the annual championship game of the *National Football League* (NFL), which was played on 2nd February 2020.

We chose the Super Bowl as the subject of our corpus, because it is not only a sporting event, but also a marketing event. Many large firms air their commercial campaigns during the televised broadcast of the Super Bowl at great expense [24]. The Super Bowl is also an entertainment event, as internationally known artists perform during the Halftime Show. There were 32.2 million interactions across Facebook, Instagram and Twitter during the Halftime Show in 2019 [1].

We retrieved all the tweets in Plymouth (UK), where the game started at 23:30. We began the retrieval two hours before the game, because this allowed us to capture the start of the televised broadcast, when a spike of Twitter activity became evident. We collected the corpus using *Tweepy* [40], an open-source, Python library for retrieving tweets in real time. Tweepy makes it easier to use the *Twitter Streaming API* by handling authentication and connection [28, 50].

Rather than retrieving tweets comprising a certain collection of hashtags, we looked for tweets containing specific keywords and phrases directly associated with the Super Bowl. Such keywords and phrases are displayed in Table 2, along with the number of tweets we collected for each of them. Note that the Twitter Streaming API is case insensitive, which guarantees the retrieval of any tweets containing the keywords in Table 2, regardless of case—for instance, the use of the keyword `superbowl` guarantees the retrieval of any tweets containing the terms `superbowl`, `SUPERBOWL`, `Superbowl`, `SuperBowl`, and any other possible case variation. The hashtag `#SuperBowl`, and all its case variations, are also retrieved by including the keyword `superbowl` in our study. Similarly, the phrase `American Football` guarantees the retrieval of any tweets including the terms `American` and `Football`, regardless of order and ignoring case.

Table 2. Number of tweets per keyword and phrase in the 2020 corpus.

Keywords and phrases	Number of tweets
<code>superbowl</code>	856,240
<code>nfl</code>	280,766
<code>football</code>	198,899
<code>touchdown</code>	37,421
<code>American Football</code>	4,478
<code>americanfootball</code>	257

The figures reported on Table 2 do not sum to give the total number of tweets available in the 2020 corpus. This is because there are many tweets which include two or more of the keywords listed in Table 2. Also, the text of some of the tweets in the 2020 corpus may not include explicit occurrences of the keywords and phrases listed in Table 2; yet, the Streaming API would provide us with such tweets if the keywords appear as part of URLs or metadata, such as user names, associated with those tweets [50]. A total of 252,678 tweets in the 2020 corpus fall into this case.

Figure 2 on page 6 shows the number of tweets captured per hour during the collection of our corpus—on average, we captured 117,311 tweets per hour. Each tweet was retrieved as a `status` object—in the context of Twitter development tools, tweets are also known as *status updates* [51]. The Streaming API provided the tweets and their corresponding metadata in Java Script Object Notation (JSON) format, and we produced a Python parser to extract the text of the tweets and other relevant information, such as the time when the tweets were published and the identifiers of the users who published those tweets. To store and manage the tweets that we collected, we uploaded them into a MySQL database, which we are using to analyse the corpus and generate statistics.

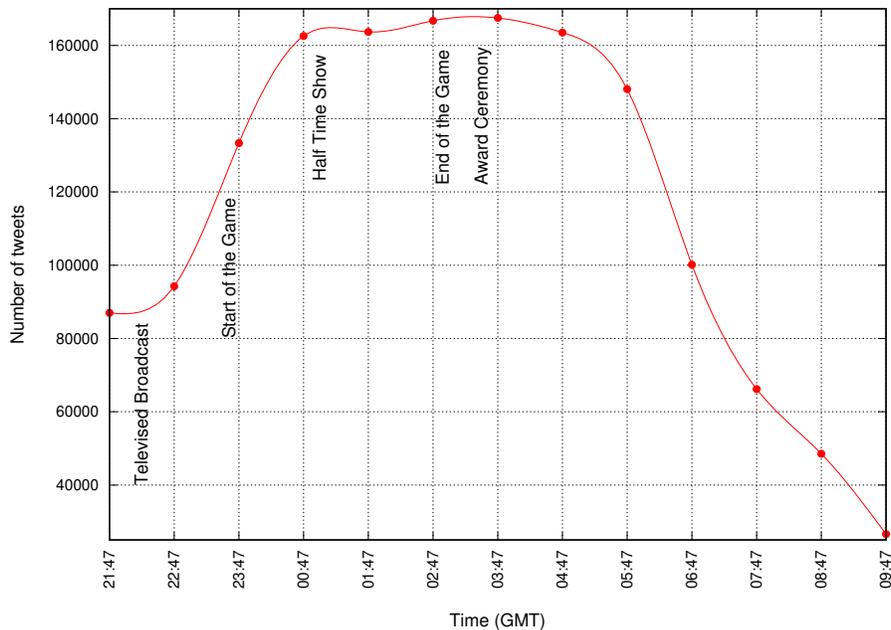


Fig. 2. Number of tweets per hour.

The Super Bowl Halftime Show began at around 01:00, which is a time when Figure 2 reaches a peak in the number of tweets collected. Plenty of people engage in conversation on Twitter during the Halftime Show. Indeed, the pinnacle of Twitter activity in the 2019 Super Bowl was logged at 01:23, when 171,000 interactions were recorded in a single minute in reaction to the Halftime Show [1]. Between 02:47 and 03:47, this is during the last quarter of the game, we gathered the largest number of tweets in the 2020 corpus: 167,492. As indicated in Figure 2, the volume of tweets started to decrease on 3rd March 2020 at 04:30, approximately—this is after the Award Ceremony had concluded.

3 Sentiment Analysis Tools

We can divide the main techniques used in sentiment analysis into *machine learning*, *lexicon-based* and *hybrid* techniques [21, 32]. Whereas machine learning uses linguistic features, the lexicon-based techniques rely on a sentiment lexicon—a collection of known and pre-compiled sentiment terms. The lexicon-based techniques are separated into *dictionary-based* and *corpus-based* approaches, which use statistical or semantic methods to determine the sentiment expressed. Hybrid techniques combine both machine learning and lexicon-based approaches with sentiment lexicons playing a critical role [25].

Traditionally, product reviews have constituted the source of data for sentiment analysis. Product reviews are important to businesses, because they can make decisions based on the analysis of the opinions about their products. However, research looking into the sentiment analysis of tweets has been widely published recently [54]: Reis *et al.* used SentiStrength to measure the negative-ness or positive-ness of news headlines [12]; O’Connor *et al.* suggested that tweets with sentiment can potentially serve as votes and substitute traditional polling [30]; and Tamersoy *et al.* explored the utilisation of VADER’s lexicon to study patterns of smoking and drinking abstinence in social media [17]. We will briefly outline below the main features of the tools chosen for our evaluation.

3.1 scikit-learn

scikit-learn is a freely-available machine learning library for the Python programming language [34]. While scikit-learn does not offer specific support on sentiment analysis, it provides all that is needed to build a classifier capable of determining the polarity of tweets.

The main reason why we chose scikit-learn over other existing alternatives is that it focuses on making machine learning available to non-specialists. Good documentation and ease of use make scikit-learn approachable and powerful. It is ideal for SMEs, which require affordable software, but it is also amply used by multinationals, such as JPMorgan, which considers scikit-learn part of its toolkit for classification and predictive analytics [43].

3.2 Sentiment140

Sentiment140 [15], formerly known as *Twitter Sentiment*, started as a student project at Stanford University, where research in sentiment analysis used to focus on large pieces of text, as opposed to tweets, which are meant to be more casual and limited to 140 characters⁴. A key contribution made by Sentiment140 at the time of its creation was the use of machine learning classifiers, rather than the then traditional lexicon-based approach.

⁴ The maximum length of a tweet used to be 140 characters. Although Twitter doubled its character length in 2017, only 1% of tweets reach the new 280-character limit, and only 12% of tweets are longer than 140 characters [20].

Given the wide range of topics discussed on Twitter, it would be too difficult to manually annotate sufficient data to train a sentiment classifier for all sorts of tweets; thus, the developers of Sentiment140 applied a technique called *distant supervision* [15], where the training data consists of tweets with emoticons. This approach was introduced by Read [38], and utilises the emoticons as “noisy” labels—for instance, :) in a tweet indicates that the tweet refers to a positive sentiment and :(indicates that the tweet expresses a negative sentiment.

Since it is relatively easy to extract several tweets containing emoticons, distant supervision is potentially a major improvement over the cost and resources that may otherwise be involved in hand-labelling training data.

3.3 SentiStrength

SentiStrength was specifically implemented to determine the strength of sentiment in informal English text, using methods to exploit the de-facto grammars and spelling styles of the informal communication that regularly takes place in social media, blogs and discussion forums [48]. Applied to *MySpace* comments, SentiStrength was able to predict positive emotion with 60.6% accuracy and negative emotion with 72.8% accuracy, both based upon numerical strength scales. SentiStrength’s prediction of positive emotion has been found to be better than general machine learning approaches [49].

To assess the results of the different tools included in this paper on the same basis, we used SentiStrength as a trinary sentiment classification tool, which means that we employed it to identify the polarity of tweets as positive, negative or neutral, though SentiStrength can also work as a binary classification tool—positive or negative.

3.4 TextBlob

TextBlob is a Python library for processing text. It offers an API to perform a number of *natural language processing* (NLP) tasks, such as noun phrase extraction, language translation and spelling correction [22]. While the most commonly known Python library for NLP is the *Natural Language Toolkit* (NLTK) [2], we favoured the selection of TextBlob in our study because it is simpler and more user-friendly than the NLTK.

With respect to sentiment analysis, TextBlob provides two options for polarity detection: **PatternAnalyzer**, which is based on the data mining *Pattern* library developed by the *Centre for Computational Linguistics and Psycholinguistics* (CLiPS) [10], and a **NaiveBayesAnalyzer** classifier, which is an NLTK classifier trained on movie reviews [36].

The default option for sentiment analysis in TextBlob is **PatternAnalyzer**, and that is precisely the option we favoured, because we are not working with movie reviews, which is the specialty of the **NaiveBayesAnalyzer** classifier. We may consider the use of the **NaiveBayesAnalyzer** classifier in the future, provided we can train it suitably for the domain of our corpus.

3.5 Treebank

Most lexicon-based sentiment analysis tools work by looking at words in isolation: giving positive points for positive words, negative points for negative words, and then summing up those points. Hence, the order of the words that compose a sentence is ignored in such tools. In contrast, Treebank, the deep learning tool for sentiment analysis developed at Stanford University, builds up a representation based on the structure of the sentences [45].

Roughly speaking, Stanford University’s deep learning model computes sentiment based on how words contribute to the meaning of longer phrases. The underlying technology is based on a new type of recursive neural network that is built on top of grammatical structures.

3.6 uClassify

uClassify was launched as a Web service in 2008, by a group of machine learning enthusiasts based in Stockholm [53]. Developers can utilise such a service to create text classifiers for various tasks, such as sentiment analysis and language detection. The uClassify sentiment classifier is trained on a corpus of 2.8 million entries comprising tweets, Amazon product evaluations and movie reviews. Hence, it can cope with both short and long texts—including tweets, Facebook statuses, blog posts and product reviews.

The uClassify API can serve a maximum of 500 requests for free on a daily basis [53]. Therefore, we would have needed several days to test uClassify with the 2020 corpus. However, the providers of this API service kindly permitted us to undertake the whole testing at once, by granting us an academic license for a limited period of time [19].

3.7 VADER

VADER, *Valence Aware Dictionary and sEntiment Reasoner*, is a rule-based tool that is specifically adapted to identify sentiments expressed in social media [17]. Using a combination of qualitative and quantitative methods, the developers of VADER built a gold-standard list of lexical features, along with their associated sentiment intensity measures. Such features are combined with consideration for five general rules, comprising grammatical and syntactical conventions for expressing and emphasising sentiment intensity.

The simplicity of VADER carries several advantages. First, it is both fast and computationally economical. Second, the lexicon and rules used by VADER are available to everyone [17]—they are not hidden within a black-box.

By exposing both the lexicon and rule-based model, VADER makes the inner workings of its sentiment analysis engine accessible—and thus, interpretable—to a broader audience beyond the scientific community.

4 Results

Figures 3 to 10 display how the polarity of the tweets in the 2020 corpus evolved per hour, according to the different tools described in Section 3.

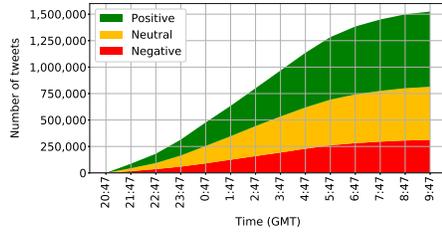


Fig. 3. scikit-learn

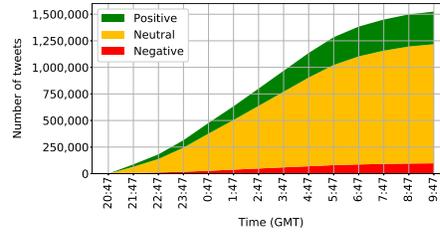


Fig. 4. Sentiment140

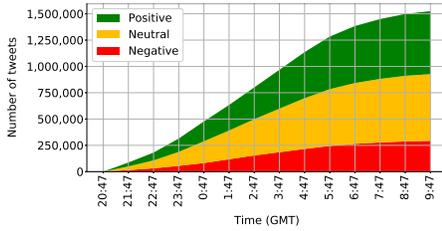


Fig. 5. SentiStrength

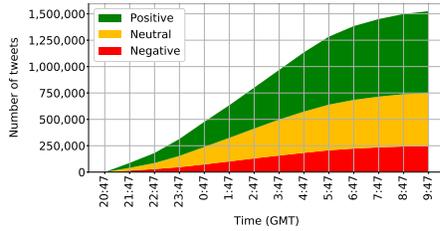


Fig. 6. TextBlob

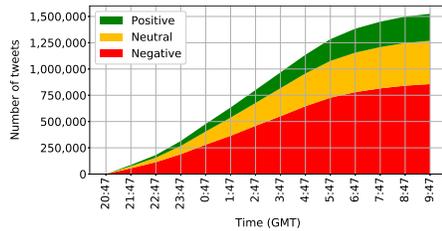


Fig. 7. Treebank

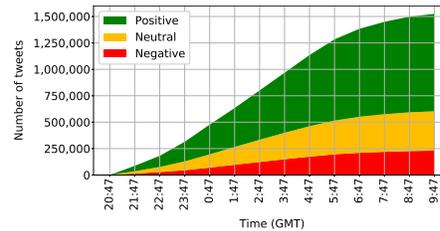


Fig. 8. uClassify

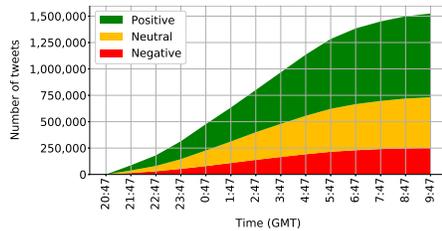


Fig. 9. VADER

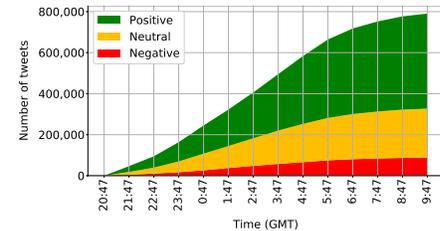


Fig. 10. Agreement

We start the analysis with Sentiment140, because it categorised as neutral more tweets than any other tool in our previous (2017) and current (2020) studies. The developers of Sentiment140 use the following litmus test to determine the polarity of a tweet: “if a tweet could ever appear as a front-page newspaper headline, or as a sentence in Wikipedia, then it is neutral” [15]. For example, the following tweet is considered neutral, because it could have been a newspaper headline, though it projects an overall negative feeling about General Motors:

RT @Finance Info Bankruptcy filing could put GM on road to profits (AP) <http://cli.gs/9ua6Sb> #Finance

The original training and test set used to develop Sentiment140 did not comprise neutral tweets—only positive and negative [15]. While not having a neutral class may have contributed to the success of the tool in previous research, it is clearly a limitation. We think this is the reason why Sentiment140 concludes the vast majority of our tweets are neutral. Indeed, Sentiment140 concludes 73% of the 2020 corpus—that is, 1,120,405 tweets—is neutral. This is so different to the results yielded by other tools that we recommend to employ Sentiment140 only if it is possible to retrain its classifier.

Retraining a classifier is precisely what we did in the case of scikit-learn. We used scikit-learn to train a linear classifier with *stochastic gradient descent* (SGD) learning [42]. To obtain suitable training and test sets, we employed a corpus of tweets gathered by Sinha *et al.* [44], which includes 290,879 tweets published by the general public at the end of NFL regular season games played in 2012. We refer to this corpus as the *2012 postgame corpus*.

Although the tweet identifiers for the 2012 postgame corpus are available at www.ark.cs.cmu.edu/football, we could not download the text of all them, as some of those tweets have already been removed from Twitter, or are no longer public. Thus, the total number of tweets we actually used to train our sentiment classifier was 100,996.

To train our classifier, we determined the polarity of the 100,996 tweets we were able to download, and then check which of them have exactly the same polarity according to two or more of the three following tools: SentiStrength, TextBlob and VADER. We chose these tools, because our analysis identified them as the ones with the greatest amount of consensus among them. SentiStrength, TextBlob and VADER agree on the classification of 52% of the 2020 corpus—that is, 790,529 tweets. Figure 10 shows how the agreement on tweet polarity changed per hour among SentiStrength, TextBlob and VADER during the retrieval of the 2020 corpus.

From the 100,996 tweets we were able to download from the 2012 postgame corpus, two or more of the chosen tools—SentiStrength, TextBlob and VADER—agree on the classification of 86,278 tweets. A total of 75% of these 86,278 tweets became our training set, whereas the remaining 25% became the test set. Our classifier achieved 90.48% accuracy, and when we used it to determine the polarity of the entire 2020 corpus, it reached a 50.23% consensus with SentiStrength, TextBlob and VADER.

Trebank returned the largest number of negative tweets in both our previous (2017) and current (2020) study. Trebank computes the sentiment based on how words compose the meaning of longer phrases [45]. For instance, Trebank recognises words such as “funny” and “witty” as positive; yet, a sentence can still be negative, regardless of the presence of positive words. Hence, the following sentence is classified as negative overall, despite the occurrences of the words “funny” and “witty” in the text:

This movie was actually neither that funny, nor super witty.

Trebank was trained using 11,855 sentences extracted from movie reviews [45]. While movie reviews are widely used in sentiment analysis [33], they form part of a completely different domain, which is beyond the scope of our corpus. Therefore, we plan to conduct further investigation on Trebank, but we will do it after retraining it with a corpus which is closer to the domain of our study: Twitter and social media in general.

Table 3 displays the consensus between any pair of tools. For example, the cell corresponding to Sentiment140 and SentiStrength indicates the percentage of tweets classified with the same polarity by both Sentiment140 and SentiStrength.

Table 3. Consensus between any pair of tools.

	scikit-learn	Sentiment140	SentiStrength	TextBlob	Trebank	uClassify	VADER
scikit-learn	--	49.01%	68.31%	81.95%	40.02%	52.25%	72.31%
Sentiment140	49.01%	--	53%	46.71%	33.42%	39.90%	48.08%
SentiStrength	68.31%	53%	--	59.70%	38.58%	49.40%	70.71%
TextBlob	81.95%	46.71%	59.70%	--	37.37%	51.96%	68.55%
Trebank	40.02%	33.42%	38.58%	37.37%	--	33.28%	40.42%
uClassify	52.25%	39.90%	49.40%	51.96%	33.28%	--	54.50%
VADER	72.31%	48.08%	70.71%	68.55%	40.42%	54.50%	--

5 Conclusions

In a landscape where little is known about the relative performance of the various sentiment analysis tools available [39], we have presented a study that aims at comparing and contrasting a selection of well-known tools. Our work, based on two different studies carried out using different corpora, reveals that the choice of sentiment analysis tool has a considerable impact on the evaluation of a corpus. Consensus among certain tools is so small that the analysis of the sentiment expressed in Twitter can be severely biased, depending on which tool is used. We suggest considering the consensus among a number of tools as a better alternative than choosing one tool and using it in isolation.

Although we started our research largely interested in supporting the needs of SMEs, our work is also of relevance to the scientific community and anyone involved in building applications using the tools discussed here.

While research on sentiment analysis continues to make progress, it remains evident that further investigation is still necessary, especially given the number of NLP problems that need to be solved first to achieve human-like performance in sentiment analysis [6]—namely, word-sense disambiguation, anaphora resolution, sarcasm detection and metaphor understanding, among others.

Acknowledgements

The authors gratefully acknowledge the free academic license provided by *uClassify* to support their experiments. We are thankful to Martin Lavelle for reading our manuscript and providing insightful comments.

References

1. Adgate, B.: What You Should Know about Super Bowl LIV Advertising and Broadcast. *Forbes* (2020)
2. Bird, S., Klein, E., Loper, E.: *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc. (2009)
3. Boyd, D., Golder, S., Lotan, G.: Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In: *Proceedings of the 43rd Hawaii International Conference on System Sciences (HICSS)*, 2010 , pp. 1–10. IEEE, Honolulu, HI (2010)
4. Cambria, E.: Affective Computing and Sentiment Analysis. *IEEE Intelligent Systems* **31**(2), 102–107 (2016)
5. Cambria, E., Das, D., Bandyopadhyay, S., Feraco, A.: Affective Computing and Sentiment Analysis. In: *A Practical Guide to Sentiment Analysis*, pp. 1–10. Springer (2017)
6. Cambria, E., Poria, S., Gelbukh, A., Thelwall, M.: Sentiment Analysis is a Big Suitcase. *IEEE Intelligent Systems* **32**(6), 74–80 (2017)
7. Collins, B.: More than 80% of SMEs Recommend Twitter for Business. *Twitter Blog* (2014)
8. Connelly, A., Kuri, V., Palomino, M.: Lack of Consensus among Sentiment Analysis Tools: A Suitability Study for SME Firms. In: *Proceedings of the 8th Language and Technology Conference*, pp. 54–8. Poznań, Poland (2017)
9. CyberEmotions: SentiStrength (2020). Available at <http://sentistrength.wlv.ac.uk/>
10. De Smedt, T., Daelemans, W.: Pattern for python. *Journal of Machine Learning Research* **13**(66), 2063–2067 (2012)
11. Devarajan, D.: Retirement of AlchemyAPI service (2017). Available at <https://www.ibm.com/cloud/blog/announcements/bye-bye-alchemyapi>
12. Dos Rieis, J.C.S., de Souza, F.B., de Melo, P.O.S.V., Prates, R.O., Kwak, H., An, J.: Breaking the News: First Impressions Matter on Online News. In: *9th International AAAI Conference on Web and Social Media*, pp. 357–366. The AAAI Press, Oxford, UK (2015)
13. European Commission: Commission Recommendation of 6 May 2003 concerning the Definition of Micro, Small and Medium-Sized Enterprises. *Official Journal of the European Union* pp. 36–41 (2003). Available at <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32003H0361>

14. Feldman, R.: Techniques and Applications for Sentiment Analysis. *Communications of the ACM* **56**(4), 82–89 (2013)
15. Go, A., Bhayani, R., Huang, L.: Twitter Sentiment Classification Using Distant Supervision. CS224N project report, Stanford **1**(12) (2009)
16. Go, Alec and Bhayani, Richa and Huang, Lei: Sentiment140 (2020). Available at <http://www.sentiment140.com/>
17. Hutto, C.J., Gilbert, E.: VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In: 8th International AAAI Conference on Weblogs and Social Media, pp. 216–225. The AAAI Press, Ann Arbor, Michigan (2014)
18. Keyvanpour, M.R., Javideh, M., Ebrahimi, M.R.: Detecting and Investigating Crime by Means of Data Mining: A General Crime Matching Framework. *Procedia Computer Science* **3**, 872–880 (2011)
19. Kågström, J.: RE: Academic Licence (Personal Communication) — E-Mail (2020)
20. Lee, K.: The Proven Ideal Length of Every Tweet, Facebook Post, and Headline Online. Fast Company pp. 1–14 (2014)
21. Liu, B.: Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies* **5**(1), 1–167 (2012)
22. Loria, S.: TextBlob Documentation. Release 0.15 **2** (2018)
23. Loria, S.: TextBlob: Simplified Text Processing (2020). Available at <https://textblob.readthedocs.io/en/dev/index.html>
24. Matheson, V.A., Baade, R.A.: Padding Required: Assessing the Economic Impact of the Super Bowl. *European Sport Management Quarterly* **6**(4), 353–374 (2006)
25. Medhat, W., Hassan, A., Korashy, H.: Sentiment Analysis Algorithms and Applications: A Survey. *Ain Shams engineering journal* **5**(4), 1093–1113 (2014)
26. Milstein, S., O’Reilly, T.: *The Twitter Book*. O’Reilly Media (2009)
27. Mohamed, A.: *The Best Software for Small Businesses (SMEs) - Essential Guide*. Computer Weekly (2009). Available at <https://www.computerweekly.com/feature/The-best-software-for-small-businesses-SMEs-Essential-Guide>. TechTarget
28. Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M.: Is the Sample Good Enough? Comparing Data from Twitter’s Streaming API with Twitter’s Firehose. In: 7th International AAAI Conference on Weblogs and Social Media, pp. 400–408. The AAAI Press, Cambridge, MA (2013)
29. Nath, S.V.: Crime Data Mining. In: *Advances and Innovations in Systems, Computing Sciences and Software Engineering*, pp. 405–409. Springer (2007)
30. O’Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In: 4th International AAAI Conference on Weblogs and Social Media, pp. 122–129. Washington, DC (2010)
31. Pak, A., Paroubek, P.: Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In: *International Conference on Language Resources and Evaluation (LREC)*, pp. 1320–1326 (2010)
32. Pang, B., Lee, L., et al.: Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval* **2**(1–2), 1–135 (2008)
33. Parkhe, V., Biswas, B.: Sentiment Analysis of Movie Reviews: Finding the Most Important Movie Aspects using Driving Factors. *Soft Computing* **20**(9), 3373–3379 (2016)
34. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011)

35. Pedregosa, et al.: scikit-learn: Machine Learning in Python (2020). Available at <https://scikit-learn.org/stable/index.html>
36. Perkins, J.: Python 3 Text Processing with NLTK 3 Cookbook. Packt Publishing Ltd. (2014)
37. Poltyn, M.: VADER Sentiment Analysis (2020). Available at <https://pypi.org/project/vader-sentiment/>
38. Read, J.: Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification. In: Proceedings of the ACL Student Research Workshop, pp. 43–48. Association for Computational Linguistics, Ann Arbor, MI (2005)
39. Ribeiro, F.N., Araújo, M., Gonçalves, P., Gonçalves, M.A., Benevenuto, F.: Sentibench—A Benchmark Comparison of State-of-the-Practice Sentiment Analysis Methods. EPJ Data Science **5**(1), 1–29 (2016)
40. Roesslein, J.: Tweepy Documentation (2020). Available at <http://docs.tweepy.org/en/v3.5.0/>
41. Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge. Speech Communication **53**(9-10), 1062–1087 (2011)
42. scikit-learn Developers: sklearn.linear_model.SGDClassifier (2019). Available at https://scikit-learn.org/0.15/modules/generated/sklearn.linear_model.SGDClassifier.html
43. scikit-learn Developers: Who is Using scikit-learn? (2019). Available at <https://scikit-learn.org/stable/testimonials/testimonials.html>
44. Sinha, S., Dyer, C., Gimpel, K., Smith, N.A.: Predicting the NFL Using Twitter. In: ECML/PKDD Workshop on Machine Learning and Data Mining for Sports Analytics. Prague, Czech Republic (2013)
45. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C.: Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank. In: Conference on Empirical Methods in Natural Language Processing, pp. 1631–1642. Association for Computational Linguistics, Seattle, WA (2013)
46. Stanford NLP Group: Sentiment Treebank (2020). Available at <https://nlp.stanford.edu/sentiment/treebank.html>
47. The Federation of Small Businesses (FSB): UK Small Business Statistics (2020). Available at <https://www.fsb.org.uk/uk-small-business-statistics.html>
48. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment Strength Detection for the Social Web. Journal of the American Society for Information Science and Technology **63**(1), 163–173 (2012)
49. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment Strength Detection in Short Informal Text. Journal of the American society for information science and technology **61**(12), 2544–2558 (2010)
50. Twitter, Inc.: Filter Realtime Tweets (2020). Available at <https://developer.twitter.com/en/docs/tweets/filter-realtime/overview>
51. Twitter, Inc.: Tweet Object (2020). Available at <https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object>
52. Twitter, Inc.: Twitter. It’s what’s happening (2020). Available at <https://twitter.com>
53. uClassify: uClassify — Free Text Classification (2020). Available at <https://www.uclassify.com/>
54. Zhou, X., Tao, X., Yong, J., Yang, Z.: Sentiment Analysis on Tweets for Social Events. In: 17th International Conference on Computer Supported Cooperative Work in Design (CSCWD), pp. 557–562. IEEE, Whistler, Canada (2013)