Lack of consensus among sentiment analysis tools: A suitability study for SME firms

Connelly, A

Lack of consensus among sentiment analysis tools: A suitability study for SME firms

Aidan Connelly [1,†,‡], **Víctor Kuri** [2,†,₸] **and Marco Palomino** [1,†,*]

[1] Big Data Group – School of Computing, Electronics and Mathematics
[2] School of Biological and Marine Sciences
[†] Plymouth University, Drake Circus, Plymouth, PL4 8AA, United Kingdom
[‡] aidan.connelly@students.plymouth.ac.uk
[₸] v.kuri@plymouth.ac.uk
[*] marco.palomino@plymouth.ac.uk

Abstract

Small and medium-sized enterprises (SMEs) are a large and integral part of the UK's economy, with over 99% of all Britain's businesses classified as small or medium. Supporting the needs of SMEs across the UK has become crucial, while the demands for software tools placed by such businesses keep growing. We are particularly interested in the case of tools for sentiment analysis, because such tools have emerged lately as a good prospect for SMEs to turn data into opportunities. We have compared five of the most well-regarded sentiment analysis tools, and have concluded that none of them is sufficiently reliable to work 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 here, while providing recommendations related to the usability of the tools in question.

Keywords: Sentiment analysis, opinion mining, Twitter, social networks, SMEs.

1. Introduction

According to the European Union and 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 (Ward and Rhodes, 2014). In the UK, small businesses accounted for 99.3% of all private sector businesses at the start of 2016 (FSB, 2017). Considering that the total employment in SMEs was 15.7 million, which equates to 60% of all private sector employment, supporting the needs of SMEs across the UK has become a key issue.

Typically, the best software to support the operation of SMEs is designed to help them to do their work while saving costs, and making their staff and processes more efficient (Mohamed, 2009). We are particularly interested in the software choices available for a specific type of application that has been gaining interest and popularity: *sentiment analysis*, the process of computationally identifying and categorising opinions expressed in a piece of text (Feldman, 2013).

The most basic task in sentiment analysis is classifying the *polarity* of a given opinion—i.e., determining whether an opinion expressed towards a particular topic or entity is *positive*, *negative* or *neutral* (Pang et al., 2002). Advanced sentiment classification may consider a variety of emotional states, such as "anger", "sadness" and "happiness", or have some discrete numeric scale into which the opinion should be categorised, like the five-star rating system used by *Amazon* (Amazon.com, 2017).

Over the past few years, several sentiment analysis tools have been developed—Ribeiro et al. (2016) claim that 7,000 articles on sentiment analysis were written up by 2016. However, despite the interest in the subject, it is still unclear which tool or method is better for identifying polarity, more convenient to adapt to different domains and purposes, or cheaper and easier to manage.

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 each specific need that businesses may encounter. Little is known about the relative performance of the various sentiment analysis tools available (Ribeiro et al., 2016); thus, comparative studies such as this one are needed. At an initial stage, our evaluation suggests that sentiment analysis can be severely biased, depending on which tool is used—even if there is agreement on the overall polarity of a corpus, major differences can be highlighted depending on the tools chosen to undertake the analysis.

We are not keen on developing new sentiment analysis tools. However, our work can be used to implement a "meta-tool" to retrieve and compare the polarity of a text according to the different tools that we have evaluated: *Sentiment140* (2017a), *SentiStrength* (2017), *Treebank* (Stanford University, 2017), *uClassify* (2017) and *VADER* (Hutto, 2017). Our source code is available at `https://github.com/AidanConnelly/SentimentConsensus`

The remainder of this paper is organised as follows: Section 2 introduces the dataset for our experiments—we gathered our own dataset to compare the tools specified above using *Twitter* (Twitter, 2017b). Section 3 describes the tools that we compared and refers to related work. Section 4 presents the results yielded by the tools that we compared and discusses our analysis. Finally, Section 5 offers our conclusions.

2. Dataset

Companies across Britain and Ireland have embraced Twitter as a powerful way to connect with their customers and grow their businesses (Collins, 2014). Twitter is now an everyday business tool for thousands of SMEs that use it for marketing, sales and customer service.

Based on the relevance of Twitter for SMEs, we have decided to use this social media platform to test a selection of sentiment analysis tools. The dataset for carrying out our tests is composed of 40,912 tweets collected at the beginning of 2017, when many people make *New Year resolutions*. Such resolutions are commonly associated with weight loss and dietary regimes. Hence, this was a good opportunity to monitor tweets related to nutritional, detox and dietary products.

Our dataset might eventually be used to perform additional studies—for example, a study on emotional response and food choice behaviour. However, we will employ it here for the evaluation of sentiment analysis tools. It is a Twitter corpus that can later lead SMEs to turn data into opportunities.

We have developed a Java-based application that interacts with the *Twitter API* (Twitter, 2017c) to retrieve public tweets. The interaction is handled by *Twitter4j* (Yamamoto, 2017). As we used the *Streaming API* (Twitter, 2017a), a stream listener retrieved the tweets that we were interested in as soon as they were published. Even though the total flow of tweets through the Streaming API is not documented, we presume that it handles up to 1% of the full firehose of tweets (140 Dev LLC, 2013).

We began the retrieval of tweets on 26 January 2017, and we ended it 20 days later—14 February 2017. We retrieved tweets in English language, exclusively. To guarantee that we gathered a good sample of tweets, a professional in the field provided the list of hashtags and phrases displayed in Table 1. Such hashtags and phrases captured conversations relevant to health and disease connected with nutritional and dietary products. Table 1 also displays the number of tweets that we collected for each hashtag and phrase.

Hashtag or phrase	No. 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

Table 1. Volume of tweets per hashtag and phrase

While some hashtags shown in Table 1 seem unintelligible to a layman, they are all sensible within the context of dietary products. For instance, the *irritable bowel syndrome*—referred to by the hashtag `#IBS`; see row 4 in Table 1—is a condition of the digestive system that is frequently mentioned in dietary conversations. Indeed, it is the third most popular hashtag in our dataset.

3. Background

Broadly speaking, there are two main approaches behind the implementation of sentiment analysis tools: *machine learning* methods that rely on supervised classification (Pang et al., 2002), and *lexicon-based* methods that employ predefined lists of words and associate each word with a specific sentiment (Tausczik and Pennebaker, 2010, Steinberger et al., 2012). Hu and Liu (2004) compiled one of such lists in 2004, and they keep updating it regularly—the current list comprises 6,800 words.

Due to its ease of use and ample reach, Twitter is rapidly changing the public discourse in society, and setting trends in topics that range from technology and entertainment to public health and politics (Kwak et al., 2010). Research looking into the sentiment analysis of tweets has been widely published. For example, Reis et al. (2015) used SentiStrength to measure the negative-ness or positive-ness of news headlines; O'Connor et al. (2010) suggested that tweets with sentiment can potentially serve as votes and substitute traditional polling; and Tamersoy et al. (2015) explored the utilisation of the VADER's lexicon (Hutto and Gilbert, 2014) to study patterns of smoking and drinking abstinence in social media.

We will briefly outline below the main features of the tools chosen for our evaluation.

3.1. Sentiment140

Sentiment140 (Go et al., 2009), formerly known as *Twitter Sentiment*, started as a class project at Stanford University (Sentiment140, 2017b), where there was already a vast amount of research in sentiment analysis, but focussed on large pieces of text, as opposed to tweets, which are meant to be more casual and limited to 140 characters. A key contribution made by Sentiment140 at the time of its creation was the use of classifiers built from machine learning algorithms, rather than the traditional lexicon-based approach.

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 tweets; thus, the developers of Sentiment140 applied a technique called *distant supervision* (Go et al., 2009), where the training data consists of tweets with emoticons. This approach was introduced by Read (2005), 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.

3.2. SentiStrength

SentiStrength was specifically implemented to determine sentiment strength from informal English text, using methods to exploit the de-facto grammars and spelling styles of the informal communication that regularly takes place in social networking websites (Thelwall et al., 2012). SentiStrength's prediction of positive emotion has been found to be better than general machine learning approaches (Thelwall et al., 2010).

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.

3.3. 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, the deep learning model for sentiment analysis developed at Stanford University, which we refer to as Treebank, builds up a representation based on sentence structure (Socher et al., 2013).

Roughly speaking, Stanford University's deep learning model computes sentiment based on how words compose 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.

3.4. uClassify

uClassify was launched as a Web service in 2008, by a group of machine learning enthusiasts based in Stockholm (uClassify, 2017). Developers can utilise this 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 (uClassify, 2017). Therefore, we would have needed 82 days to test uClassify with our dataset. 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 (Kågström, 2017).

3.5. VADER

VADER—*Valence Aware Dictionary and sEntiment Reasoner*—is a rule-based tool that is specifically adapted to identify sentiments expressed in social media (Hutto and Gilbert, 2014). 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 accessible to anyone (Hutto and Gilbert, 2014)—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

The polarity results for the different sentiment analysis tools chosen in this study are presented in Table 2. It should be observed that the results yielded by the five tools are so dissimilar that we cannot trust in any one of them without further investigation.

While Sentiment140 considers 1% of the dataset as negative, Treebank considers 70% of it as negative. We cannot rely on such disparate results. Also note that Treebank and uClassify provide similar figures for the number of neutral tweets, but their disagreement on the classification of positive and negative tweets is enormous.

	Positive	Negative	Neutrals
Sentiment140	9,285	439	31,188
SentiStrength	16,224	5,684	19,004
Treebank	5,739	27,505	7,668
uClassify	31,323	2,396	7,193
VADER	4,548	274	36,090

Table 2. Polarity per tool

A possible alternative to selecting an individual tool consists of employing more than one tool simultaneously, and rely only on the classification of tweets for which all the tools reach an agreement. To explore this alternative, we calculated the consensus among the five tools evaluated. Table 3 shows the consensus, which is very small—1,559 tweets, or 3.81% of the dataset. It should be observed that there is 38 times more consensus on neutral tweets—849—than on negative ones—22.

	Number of tweets	Proportion (%)
Positive	688	1.68
Negative	22	0.05
Neutrals	849	2.08
Discrepancy	39,353	96.19
Total	40,912	100

Table 3. Total consensus and discrepancy

We have also identified the cases where 4 or less tools agreed on the polarity of the tweets, and this is presented in Table 4 and Table 5. Note that 4 out of the 5 tools agreed on the polarity of 8,082 tweets—19.75% of the dataset—see column 7 in Table 4. Also, 3 out of the 5 tools agreed on the polarity of 21,484 tweets—52.51% of the dataset—see column 8 in Table 4. Additionally, Table 4 shows which specific tool disagrees with the rest—see columns 2-6 in Table 4.

We have also calculated the consensus between any pair of tools. Table 5 displays these calculations: the row corresponding to SentiStrength and the column corresponding to VADER, in the section marked as "Positive Consensus", shows the number of tweets classified as positive by both SentiStrength and VADER. Table 6 displays examples of tweets that are part of the consensus in each category: positive, negative and neutral.

In terms of usability, VADER seems the "friendliest" tool to use—a few PIP commands are enough to configure it—whereas Treebank is both the most complicated tool to use and the slowest one to perform. Treebank requires 6,586 seconds—i.e., 1 hour, 49 minutes and 46 seconds—to compute the polarity of the entire dataset on an AMD Athlon X4 860K processor at 3.7GHz. Comparatively, the fastest tool, SentiStrength, requires only 9.74 seconds to perform the same task using the same equipment. The rest of the tools performed as follows: Sentiment140, 322.11 seconds; uClassify, 191.92 seconds; and VADER, 18.47 seconds—all values are the average after 40 executions.

5. Conclusions

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 (IBM, 2017a), which is now a core component of IBM's *Watson Developer Cloud* (IBM, 2017b). However, most SMEs would find it unreasonable to invest significantly in such an activity. Hence, we have produced this evaluation.

The outcomes of the five tools that we evaluated are so contrasting and diverging from each other that we cannot trust in any one of them without further investigation. While we suggest considering the consensus among various tools as a better alternative than choosing one and using it in isolation, we emphasise that any analysis of the sentiment expressed in social media can be severely biased, depending on which tools are used.

	Tool that the does not agree with the other 4						
	Sentiment140	**SentiStrength**	**Treebank**	**uClassify**	**VADER**	**Consensus 4/5**	**Consensus 3/5**
Negative	48	2	1	12	104	167	704
Positive	532	48	1,173	48	1,173	2,974	5,447
Neutral	103	575	2,049	2,147	67	4,941	15,333
Total	683	625	3,223	2,207	1,344	8,082	21,484

Table 4. Consensus for 3-4 out of 5 tools

Positive Consensus					
	Sentiment140	**SentiStrength**	**Treebank**	**uClassify**	**VADER**
Sentiment140	9,285	16,469	5,955	5,974	28,654
SentiStrength	16,469	16,224	3,651	3,461	17,916
Treebank	5,955	3,651	5,739	1,783	6,249
uClassify	5,974	3,461	1,783	31,323	6,700
VADER	28,654	17,916	6,249	6,700	4,548
Negative Consensus					
	Sentiment140	**SentiStrength**	**Treebank**	**uClassify**	**VADER**
Sentiment140	439	248	378	211	41
SentiStrength	248	5,684	4,318	782	218
Treebank	378	4,318	27,505	1,765	169
uClassify	211	782	1,765	2,396	105
VADER	41	218	169	105	274
Neutral Consensus					
	Sentiment140	**SentiStrength**	**Treebank**	**uClassify**	**VADER**
Sentiment140	31,188	6,352	2,309	8,038	2,218
SentiStrength	6,352	19,004	4,040	13,855	3,439
Treebank	2,309	4,040	7,668	5,043	1,536
uClassify	8,038	13,855	5,043	7,193	4,058
VADER	2,218	3,439	1,536	4,058	36,090

Table 5. Consensus for any pair of tools

Tweet	**Polarity**
Awesome! Love the Instagram post, you are very dedicated;-) which indicates success! Keep us posted on your...	Positive
This is awesome!	Positive
When your stomach hurts so badly that you just resign yourself to impending death: Yup Im dying. Hurts so bad	Negative
"IBS is seriously so draining. Im either hungry, bloated, or in pain. ???#ibs"	Negative
Diabetics experience #IBS like symptoms. Find out how they are connected and how you can manage	Neutral
"Check out Sugar Alternatives: Lemons & Limes #kitchology, #foodallergy"	Neutral

Table 6. Examples of tweets that are part of the consensus

References

140 DEV LLC. 2013. *Aggregating tweets: Search API vs. Streaming API* [Online]. Twitter API Programming Tips, Tutorials, Source Code Libraries and Consulting. Available: http://140dev.com/twitter-api-programming-tutorials/aggregating-tweets-search-api-vs-streaming-api/ [Accessed 2017].

AMAZON.COM. 2017. *Receiving Amazon.co.uk Feedback from Buyers* [Online]. Amazon.com, Inc.

Available: https://www.amazon.co.uk/gp/help/customer/display.html?nodeId=13841791 [Accessed 2017].

COLLINS, B. 2014. *More than 80% of SMEs recommend Twitter for business* [Online]. Available: https://blog.twitter.com/marketing/en_gb/a/en-gb/2014/more-than-80-of-smes-recommend-twitter-for-business.html [Accessed 2017].

FELDMAN, R. 2013. Techniques and applications for sentiment analysis. *Communications of the ACM,* 56**,** 82-89.

FSB. 2017. *UK Small Business Statistics* [Online]. Blackpool, UK: National Federation of Self Employed & Small Businesses Limited. Available: https://www.fsb.org.uk/media-centre/small-business-statistics [Accessed 2017].

GO, A., BHAYANI, R. & HUANG, L. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford,* 1**,** 12.

HU, M. & LIU, B. 2004. Mining and Summarizing Customer Reviews. *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining.* New York, NY: ACM.

HUTTO, C. 2017. *VADER - Sentiment Analysis* [Online]. GitHub, Inc. Available: https://github.com/cjhutto/vaderSentiment [Accessed 2017].

HUTTO, C. J. & GILBERT, E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International AAAI Conference on Weblogs and Social Media, 2014 Ann Arbor, MI. 216-225.

KÅGSTRÖM, J. 25 July 2017. *RE: Academic Licence (Personal Communication) – E-Mail*.

KWAK, H., LEE, C., PARK, H. & MOON, S. What is Twitter, a social network or a news media? Proceedings of the International Conference on World Wide Web, 2010 Raleigh, NC. ACM, 591-600.

MOHAMED, A. 2009. The best software for small businesses (SMEs) - Essential Guide. *Computer Weekly*.

O'CONNOR, B., BALASUBRAMANYAN, R., ROUTLEDGE, B. R. & SMITH, N. A. From tweets to polls: Linking text sentiment to public opinion time series. Proceedings of the International AAAI Conference on Weblogs and Social Media, 2010 Washington, DC. 1-2.

PANG, B., LEE, L. & VAITHYANATHAN, S. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2002 Philadelphia, PA. Association for Computational Linguistics, 79-86.

READ, J. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. Proceedings of the ACL Student Research Workshop, 2005 Ann Arbor, MI. Association for Computational Linguistics, 43-48.

REIS, J., BENEVENUTO, F., DE MELO, P. V., PRATES, R., KWAK, H. & AN, J. Breaking the news: First impressions matter on online news. Proceedings of the International Conference on Weblogs and Social Media, 2015 Oxford, UK.

RIBEIRO, F. N., ARAÚJO, M., GONÇALVES, P., GONÇALVES, M. A. & BENEVENUTO, F. 2016. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science,* 5**,** 1-29.

SENTIMENT140. 2017a. *Sentiment140 - A Twitter Sentiment Analysis Tool* [Online]. Available: http://www.sentiment140.com/ [Accessed 2017].

SENTIMENT140. 2017b. *Sentiment140 - For Academics* [Online]. Available: http://help.sentiment140.com/for-students [Accessed 2017].

SENTISTRENGTH. 2017. *SentiStrength - Sentiment strength detection in short texts* [Online]. Available: http://sentistrength.wlv.ac.uk/ [Accessed 2017].

SOCHER, R., PERELYGIN, A., WU, J., CHUANG, J., MANNING, C. D., NG, A. & POTTS, C. Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of the Conference on Empirical Methods in Natural Language Processing, October 2013 Seattle, WA. The Association for Computational Linguistics, 1631-1642.

STANFORD UNIVERSITY. 2017. *The Stanford Natural Language Processing Group* [Online]. Available: https://nlp.stanford.edu/ [Accessed 2017].

STEINBERGER, J., EBRAHIM, M., EHRMANN, M., HURRIYETOGLU, A., KABADJOV, M., LENKOVA, P., STEINBERGER, R., TANEV, H., VÁZQUEZ, S. & ZAVARELLA, V. 2012. Creating sentiment dictionaries via triangulation. *Decision Support Systems,* 53**,** 689-694.

TAMERSOY, A., DE CHOUDHURY, M. & CHAU, D. H. Characterizing smoking and drinking abstinence from social media. Proceedings of the ACM Conference on Hypertext & Social Media, 2015 Cyprus. ACM, 139-148.

TAUSCZIK, Y. R. & PENNEBAKER, J. W. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology,* 29**,** 24-54.

THELWALL, M., BUCKLEY, K. & PALTOGLOU, G. 2012. Sentiment strength detection for the social web. *Journal of the Association for Information Science and Technology,* 63**,** 163-173.

THELWALL, M., BUCKLEY, K., PALTOGLOU, G., CAI, D. & KAPPAS, A. 2010. Sentiment strength detection in short informal text. *Journal of the Association for Information Science and Technology,* 61**,** 2544-2558.

TWITTER. 2017a. *Streaming APIs* [Online]. Twitter, Inc. Available: https://dev.twitter.com/streaming/overview [Accessed 2017].

TWITTER. 2017b. *Twitter (@Twitter)* [Online]. Available: https://twitter.com/twitter?lang=en [Accessed 2017].

TWITTER. 2017c. *Twitter Developer Documentation - API Overview* [Online]. Available: https://dev.twitter.com/overview/api [Accessed 2017].

UCLASSIFY. 2017. *uClassify - Free text classification* [Online]. Available: https://www.uclassify.com/ [Accessed 2017].

WARD, M. & RHODES, C. 2014. Small Businesses and the UK Economy. House of Commons Library.

YAMAMOTO, Y. 2017. *Twitter4J - An unofficial Java library for the Twitter API* [Online]. Available: http://twitter4j.org/en/ [Accessed 2017].