Gender Bias in Sentiment Analysis[1]

Purpose: To test if there are biases in lexical sentiment analysis accuracy between reviews authored by males and females.
Design: This paper uses datasets of TripAdvisor reviews of hotels and restaurants in the UK written by UK residents to contrast the accuracy of lexical sentiment analysis for males and females.
Findings: Male sentiment is harder to detect because it is less explicit. There was no evidence that this problem could be solved by gender-specific lexical sentiment analysis.
Research limitations: Only one lexical sentiment analysis algorithm was used.
Practical implications: Care should be taken when drawing conclusions about gender differences from automatic sentiment analysis results. When comparing opinions for product aspects that appeal differently to men and women, female sentiments are likely to be overrepresented, biasing the results.
Originality/value: This is the first evidence that lexical sentiment analysis is less able to detect the opinions of one gender than another.

Keywords: Sentiment analysis; opinion mining; social media; online customer relations management.

Introduction

Sentiment analysis is the computer-based estimation of the sentiment expressed in text, such as its overall polarity, the range of emotions expressed, or the strengths of any opinions. It is widely used within marketing and customer relations management through the online monitoring of customer opinions towards products and services expressed in social media or review sites (Pekar & Ou, 2008; Schweidel & Moe, 2014; Tirunillai & Tellis, 2014). Because of this, social media monitoring has been a standard business technique for over half a decade (Hofer-Shall, 2010). It takes advantage of the public availability of opinionated texts and fast, accurate software for detecting opinions. It can also be used to assess the impact of business interventions in the social web (Homburg, Ehm, & Artz, 2015).

Sentiment analysis is typically used as a black box solution by marketers who see the results of the algorithm used to classify sentiment but are not interested in its details. For example, they may find that 45% of comments about product A are positive in comparison to 25% for product B, concluding that product A is more favourably viewed. This may be misleading if there is bias in the data or the sentiment analysis algorithm. For example, if product B's admirers are older and less likely to post to the social web then their opinions would be underrepresented. The importance of representativeness is recognised for survey-based research (e.g., Gronholdt, Martensen, & Kristensen, 2000) and it is equally important for big data analyses. Moreover, no automatic sentiment analysis system is perfect and it is possible that the sentiment in posts about product A are harder to detect, introducing a hidden (to the marketer) source of bias. This bias may occur if the positive aspects of product A are difficult to describe explicitly or if its admirers are from a group that express sentiment less directly, the case that is considered here for gender. For example, a sentiment-based comparison of smartphones (Kim, Dwivedi, Zhang, & Jeong, 2016) might give gender biased results if the system is better at identifying sentiment from one gender

[1] Thelwall, M. (in press). Gender bias in sentiment analysis. *Online Information Review*. doi: 10.1108/OIR-05-2017-0139

than from another, and a system that detects sentiment to help select good ideas (Lee & Suh, 2016) might have a bias towards the opinions of one gender.

On Twitter, some emotion-related terms, such as love and haha, are disproportionately used by one gender. The same is true for some other words and linguistic features, such as exclamation marks (Burger, Henderson, Kim, & Zarrella, 2011; Volkova & Yarowsky, 2014). More generally, there are gender differences in the extent to which sentiment is expressed in the social web (Bagić Babac & Podobnik, 2016; Rangel & Rosso, 2016; Thelwall, Wilkinson, & Uppal, 2010; Volkova & Yoram, 2015), perhaps echoing common stylistic differences (e.g., Koppel, Argamon, & Shimoni, 2002; Thelwall, 2016). It is not clear whether the differences are at the sentiment level or the linguistic levels, however, because the differences could be stylistic or substantive. Thus, sentiment analysis algorithms may perform differently for males than for females. If true, then sentiment analysis is currently providing misleading information to businesses about products and services that are viewed differently by males and females. There is therefore an urgent need to test for the existence of gender biases in the accuracy of sentiment analysis algorithms.

This paper assesses the influence of gender on the accuracy of sentiment analysis using case studies of TripAdvisor hotel and restaurant reviews, chosen as product types that are similarly important for both males and females, without obvious gender differences. There are two main sentiment analysis strategies (Liu, 2012; Pang & Lee, 2008). Machine learning techniques typically split reviews into collections of words or phrases and then train an algorithm to detect patterns of association with sentiment (e.g., Turney, 2002). In contrast, lexical algorithms exploit a list of sentiment bearing terms and apply a set of rules to deduce the sentiment of a text primarily from occurrences of these words (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). This paper uses the lexical approach as the harder test case because its limit to a set lexicon makes it less likely to be influenced by gender-specific terms. There are many lexical sentiment analysis algorithms with different lexicons and combinations of linguistic rules. Some also use part of speech tagging for sentiment word sense disambiguation (Baccianella, Esuli, & Sebastiani, 2010) and to treat some parts of speech, such as adjectives, differently (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). The algorithm used here is SentiStrength (Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010), which is a widely used lexical algorithm and incorporates a machine learning element that can be used to adjust it to specific tasks, such as for gender-specific sentiment analysis (see below).

This article also assesses whether it is possible to improve lexical sentiment analysis accuracy by incorporating gender information. This is a similar rationale to that of a gender-aware word sense disambiguation program for blog posts (Mihalcea & Garimella, 2016). A previous study has indirectly incorporated author gender into subjectivity and sentiment polarity detection for Twitter in English, Spanish and Russian. Removing terms, such as *weakness*, that tend to have different polarities for males and females, was shown to improve machine learning accuracy for sentiment analyses of 2,000 tweets for each language (Volkova, Wilson, & Yarowsky, 2013). Since the Twitter corpora were unrestricted by topic, it is possible that some words were used by males for one topic but by females for another (e.g., dogfighting) so that the improved performance could be due to differing topics rather than methods of expressing sentiment.

Sentiment strength detection with SentiStrength

The sentiment strength detection task is to detect both the overall sentiment polarity of a text and its strength, or to annotate it with a numerical rating. As mentioned above, a basic machine learning approach first converts texts into collections of constituent words and short phrases, known as *features*. Human sentiment classifications of these texts are then used to train a machine learning algorithm to understand which features associate with sentiment. For example, an algorithm might learn from the human classified examples that the bigram "we love" is a good indicator of strong positive sentiment. In contrast, a lexical sentiment analysis algorithm incorporates a lexicon of sentiment words with their typical sentiment strengths and polarities. When fed a new text, the program uses the presence of sentiment words in its lexicon to estimate the likely sentiment, with the aid of additional linguistic information and rules (Thelwall, Buckley, & Paltoglou, 2012).

The lexical software SentiStrength has a lexicon of 2,846 sentiment words and word stems with human-annotated sentiment strengths and polarities on a scale of --5 (very strong negativity) to 5 (very strong positivity). The number 0 is not used and 1 and -1 indicate the absence of positive and negative sentiment, respectively. The sentiment strength scores for the terms in the lexicon were initially assigned by a human coder but subsequently optimised using machine learning (see the methods section for relevant details) with training sets from various social web sites (e.g., Twitter, YouTube, BBC News discussions) and, finally, TripAdvisor (ignoring reviewer gender) to optimise the lexicon for detecting the strength of sentiment in TripAdvisor reviews (Appendix, Figure 4). SentiStrength allocates each sentence the highest score of any positive word and the highest score of any negative word (i.e., a positive sentiment score in the range 1 to 5, and a negative sentiment score in the range -1 to -5) and then adds the results to give an overall sentiment strength and polarity output in the range -4 to 4. Before adding the positive and negative scores, it applies linguistic rules (Thelwall, Buckley, & Paltoglou, 2012). For instance, the negation rule reverses the polarity of a term when preceded by a negator, such as *not*, *don't* and *wouldn't*. Similarly, preceding booster words (e.g., very, just) either strengthen or weaken the strength of the subsequent word. For example, the phrase, "Really wonderful food but boring show" scores +3 for *wonderful*, which is boosted to +4 by *really*. It also scores -2 for *boring*, giving a combined score of +2 on the -4 to +4 scale.

Research questions

The research questions are posed in general but will be assessed using two specific test cases: hotel reviews and restaurant reviews. The focus is on narrow topics, such as these, on the assumption that gender differences will be more substantial for broad topics due to gender differences in preferences for different aspects of the topics.

1. Are there author gender differences in the accuracy of lexical sentiment strength detection on narrow topics?
2. Can the accuracy of lexical sentiment detection for narrow topics be increased by optimising separately for male and female authors?

Methods

The research design was to gather a large collection of product reviews with associated ratings and author genders and then to assess and subsequently optimise for accuracy a sentiment analysis program on an appropriate subsample for (a) males, (b), females and (c) all users. The SentiStrength software is appropriate for this because this is designed to detect sentiment strength rather than just polarity. It is lexical (with its own dictionary: Thelwall, Buckley, & Paltoglou, 2012) with machine learning features and this combination is needed to be able to optimise separately by gender and then manually interpret the results (i.e., the gender differences in the dictionaries).

TripAdvisor reviews was chosen as the data source because contains many hotel and restaurant reviews and allows data harvesting. Its reviews seem to be mostly high quality and relatively spam-free. The UK TripAdvisor website www.tripadvisor.co.uk contains only reviews in English about UK locations. The UK was chosen rather than the USA or any other English-speaking nation because there are international differences in versions of English and patterns of informal use and so, other factors being equal, it is preferable to analyse cultural issues from countries that an author is familiar with. In this case, the use of the UK will help with error analysis and interpreting any gender difference results at the level of individual words.

TripAdvisor provides a sitemap with an apparently comprehensive list of all pages on its UK site https://www.tripadvisor.co.uk/sitemap/2/en_UK/sitemap_en_UK_index.xml. This sitemap was downloaded on 6 February 2017 and its URLs extracted to form a complete list. From this list the pages with user reviews were selected using the URL filter *ShowUserReviews*. Since there were many pages, they were additionally filtered to England (e.g., excluding Scotland, Wales, and Northern Ireland) with the URL filter *_England.html to obtain a slightly more homogeneous set. This resulted in 13,234,039 user review page URLs. This list was partitioned into 20 equal sized files using a random number generator and one file was selected as the development set (for use prior to the main experiments) and another as the evaluation set. All these filtering stages were conducted using the free software Webometric Analyst.

The www.tripadvisor.co.uk website allows crawling with its robots.txt file (Thelwall & Stuart, 2006) and the free web crawler SocSciBot was used to download the two random sets of England review URLs during February and March 2017 at a rate of one URL per second, obeying robots.txt commands.

Once the downloading was complete, code was added to the free software Mozdeh to extract all user reviews from each downloaded page. For each review, the text, the author username, their self-declared geographic location and overall rating was extracted. The nature of the attraction (e.g., hotel, restaurant) and hotel star ratings were also extracted from each page. Although TripAdvisor does not enforce a standard method of geographic location on its users, inspection of the data showed that most seemed to declare a country or major town. UK-based reviewers were therefore identified on the basis that they mentioned the UK, a constituent country, or one of the ten largest cities. This may generate a small percentage of false matches (e.g., from London, Ontario) but overall produces a large and mostly reliable dataset. A manual examination of 1000 matches of the queries found no obvious false matches. For example, of the 165 people giving London as their location (i.e., the least reliable case), all except 29 also included the UK, United Kingdom or England and these 29 gave no additional location information and so could not be checked further (other than one including UK in her username). It seems likely that all of these were from the UK London. Birmingham is another city with a large overseas equivalent (Alabama) but all people from Birmingham either stated the UK or had reviewed a location close to the UK Birmingham and so were almost certainly from the UK. Thus, the overall geographic location error rate may be below 0.1%. Duplicate reviews were also removed at this stage. This could occur if the author accidentally posted twice or if the review appeared on multiple pages.

Although users can declare their gender in TripAdvisor, this information is not displayed with their reviews but only in their profile page. Profiles were not downloaded because this would have added substantially to the TripAdvisor crawl with a small benefit. Instead, US census 1990 information was used listing the genders of the most popular 10,000 first names in the country. The 4772 names in this list that had a gender orientation of at least 90% and at least four letters were selected as reliably indicating either male or female. Usernames were gendered by comparing the first name to these two lists for matches. Usernames that were single terms were split into two if using camel case (e.g., SarahSingh) and the first segment taken as the given name. Genders were not assigned to the 70% of reviewers that did not get a match from this process.

Code was added to the free Windows version of SentiStrength to generate gender-specific (using the methods above) and attraction-specific balanced subsets of the data (the Convert menu). TripAdvisor ratings are on a five-point decile scale (10, 20, 30, 40, 50) and these were converted to SentiStrength's scale system (-4, -3, -2, -1, 0, 1, 2, 3, 4) as follows; (10 -> -4, 20 -> -2, 30 -> 0, 40 -> 2, 50 -> 4). This does not use four of the SentiStrength output values (-3, -1, 1, 3) and so SentiStrength gives slightly more fine-grained results than reviewers.

The two largest genres were selected for analysis: hotels and restaurants. Hotels are given star ratings in TripAdvisor and homogeneous hotel subsets were built from the three most numerous ratings (2, 3 and 4 star). For each of the four sets, two random equal ratings samples were created, one for male reviewers and one for female reviewers. Here, an equal ratings random sample means that all five ratings (10, 20, 30, 40, 50) are equally represented. To achieve this, the number of reviews was calculated for each rating and all reviews selected for this. For the remaining four ratings, a random sample was drawn with the same sample size as the smallest set (all using the SentiStrength Convert menu). These data sets were then used to compare the accuracy of SentiStrength separately for male-authored and female-authored reviews (Table 1).

Table 1. Sample sizes of the data sets used. Datasets have equal numbers of ratings of 10, 20, 30, 40, and 50. Each dataset has at most one review per TripAdvisor user (hence the Hotels – all category is not larger than the sum of the individual hotels category).

Topic	**Male-authored reviews**	**Female-authored reviews**
Hotels - 2.0 star	1695	1695
Hotels - 3.0 star	6700	6700
Hotels - 4.0 star	6215	6215
Hotels - all	11900	11900
Restaurant	21255	21255

For the second research question, the hill climbing dictionary optimisation routine in the Java version of SentiStrength (Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010) was used to optimise the sentiment term weights for each single gender set. It selected each term in the sentiment dictionary at random and then assessed whether the accuracy of the classifications on the input set would be improved if the sentiment weight was increased or decreased, keeping changes that make an overall improvement and an individual improvement on at least two different texts. This process was repeated for all terms in the dictionary until no changes had been made after a full round. This algorithm was used in 10-fold cross-validation mode to estimate the performance change caused by the optimisation on each set individually. Since the optimisation algorithm includes an element of randomness for the order, 30 repetitions were used to generate 95% confidence intervals. After each training phase, the optimised algorithm was evaluated on both the hidden fold from the training gender and on the full set for the non-training gender. For example, for the female-authored restaurant reviews, the algorithm was trained 30x10=300 times, each using 90% of the female-authored restaurant reviews. It was also evaluated 300 times on the remaining 10% of the female-authored restaurant reviews and 100% of the male-authored restaurant reviews (since they were not used for training).

The performance of the algorithm was estimated using mean absolute deviation (MAD) scores. These give the average of the absolute differences between the user ratings and SentiStrength predictions. MAD is better than the traditional precision and recall formulae for scalar data. The primary analysis method is therefore to compare the MAD scores between the different versions of SentiStrength.

Results

The accuracy of the TripAdvisor variant of SentiStrength is statistically significantly lower for males than for females on all five datasets (Figure 1: higher bars for MM and FM than for MF and FF within each of the five sets). This gives strong evidence of gender bias, answering the first research question. In the largest case (Hotels 3.0) the difference in MAD is 0.09, or 4.5% of the difference between two TripAdvisor ratings levels (e.g., 40 vs. 50), which is not substantial, however.

Training on the same-gender data set only statistically improves accuracy for female-authored Hotels-All reviews in comparison to training on the opposite gender (Figure 1: overlapping confidence intervals between MM and FM and between MF and FF within each of the five sets, except Hotels-All MF/FF). In the statistically significant case, the MAD increase is about 0.01, or 0.5% of the difference between two TripAdvisor ratings categories, which is minor. Given that statistically spurious results are more likely when multiple comparisons are performed and the same gender training is not always an improvement (irrespective of statistical significance) in the different tests, this gives a negative answer to the second research question: the results do not suggest that same gender training can improve lexical sentiment analysis accuracy, at least for SentiStrength and UK English TripAdvisor reviews of hotels and restaurants.

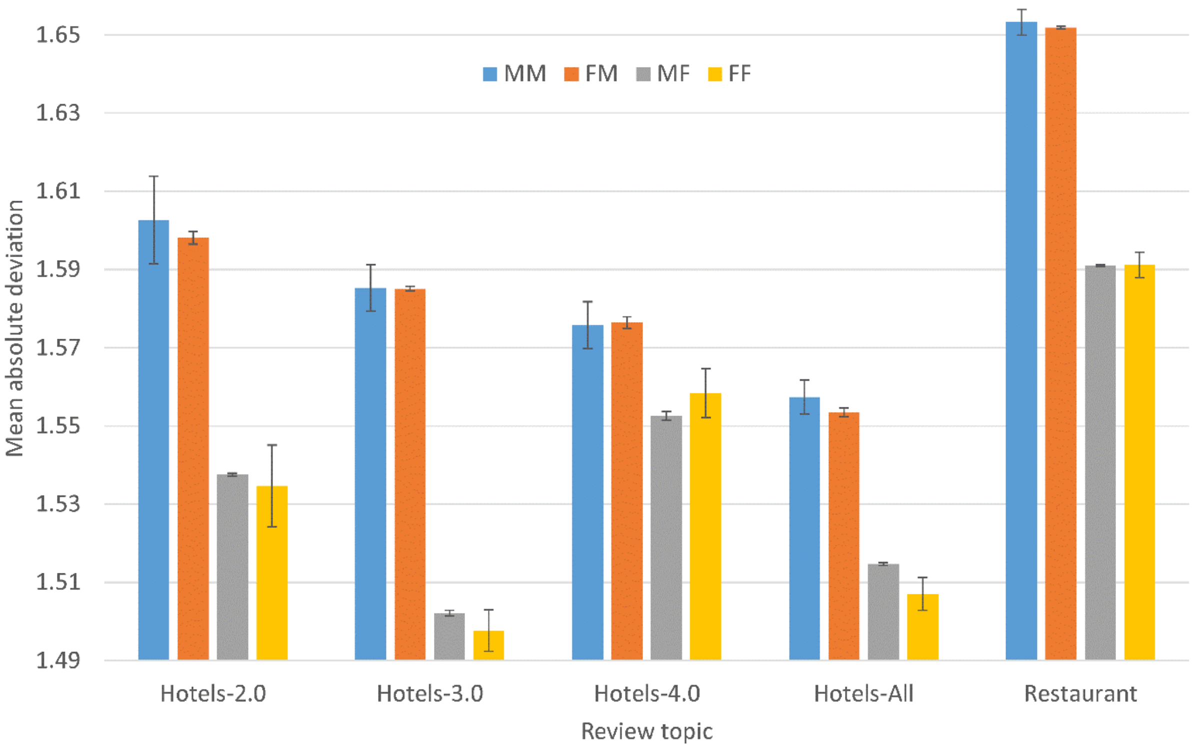

Figure 1. Mean absolute deviation (lower values indicate higher accuracy) for SentiStrength estimates of review sentiment on a scale of -4 to 4, compared to reviewer ratings for sample equalised collections of hotel and restaurant reviews, after sentiment term weight optimisation. In the legend, the first gender is the training set and the second letter is the evaluation set. Confidence intervals are wider for the same gender columns because the evaluation sets are smaller (10% rather than 100%). Non-overlapping confidence intervals indicate statistical significance at the 95% level and statistical significance is also possible for small overlaps.

Reasons for gender differences

There are different potential explanations for the higher accuracy for female-authored reviews. Male authored reviews might be more balanced, by attempting to be professional with both positive and negative points in all reviews. This would make the prediction of sentiment more difficult. Alternatively, males might use more sarcasm, figurative language or more obscure terminology than females, to inject either humour or a sense of professional competence into reviews. Since sentiment is difficult to detect in figurative language is difficult to detect, this would cause accuracy differences.

To explore gender differences for individual ratings levels, male-authored reviews, were compared to female-authored reviews at each score level (Figures 2 and 3). For both genders and for both hotels and restaurants, SentiStrength substantially underestimated the number of negative reviews. Within these results, the female columns tend to be higher than the male columns for sentiment at either extreme (very positive or very negative) whereas the reverse tends to be true for milder sentiment. Given that SentiStrength detects linguistic expressions of sentiment, this is consistent with two different explanations: females using more explicit positive language than males, or avoiding negative terms in strongly positive reviews and positive terms in strongly negative reviews.

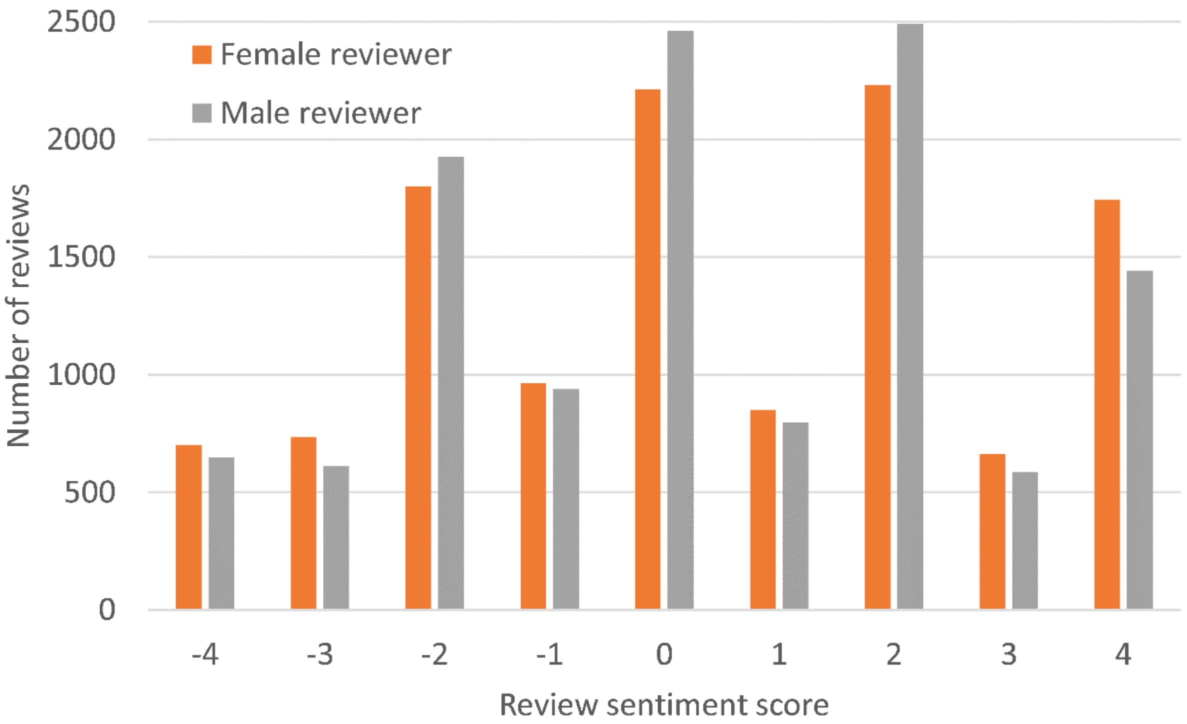

Figure 2. SentiStrength estimated sentiment of reviews (before machine learning) for the male and female authored reviews in the Hotels-All data set. The data sets are balanced so the correct scores for each gender would be identical column heights for all the male bars and all the female bars at each of the levels -4, -2, 0, 2, 4.

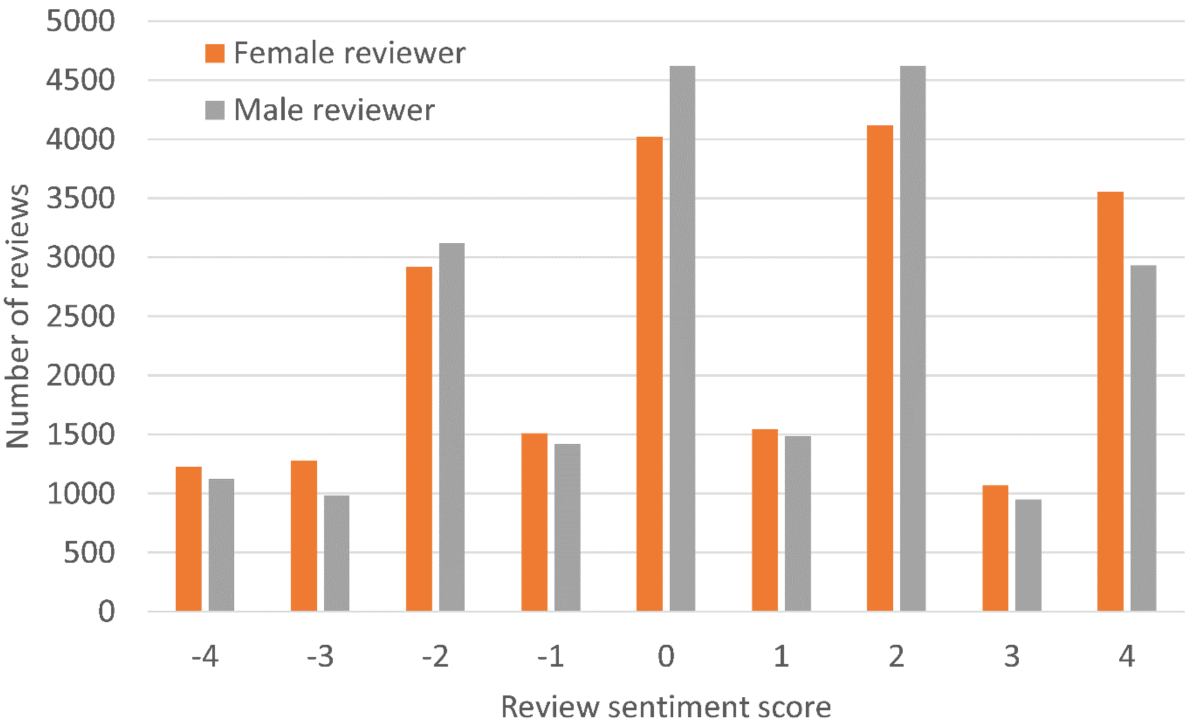

Figure 3. As Figure 2 for the Restaurant data set.

SentiStrength can fail to detect positive and negative expressions if they do not include any sentiment terms. This is the reason for the lower levels of extreme positive and negative sentiment. Examples of phrases in reviews of 50-rated hotels that implied strong approval but that were not detected by the sentiment analysis algorithm include, “the main was out of this world”, “Will going back very soon”, “the view was unreal”, “staff were so informative”, “would go back time and time again”, “shocked by the variety”, “food and service were first class”, “so much more character”. Some reviews also had typos in the sentiment terms, preventing them from being detected.

A word frequency analysis was used to identify the main differences between terms in male and female authored reviews with the highest and lowest ratings (50 and 10, corresponding to 4 and -4, respectively). For the highest rated hotel reviews, terms that occurred in a higher proportion of male-authored than female-authored 50-rated reviews were ranked using difference between proportions z values and the top 20 terms were manually examined for patterns. This was repeated for the lowest rated reviews, and again after switching gender roles and for the restaurant reviews (Table 2, 3).

Table 2. Words that occurred disproportionately often in the 50-rated reviews of one gender in comparison to the other gender. The percentage of reviews containing the term for the majority and other gender is shown in brackets.

Rank	Hotels, female	Restaurants, female	Hotels, male	Restaurants, male
1	lovely (31%/17%)	lovely (31%/16%)	wife (9%/1%)	wife (8%/1%)
2	husband (7%/1%)	husband (7%/1%)	beer (6%/2%)	beer (5%/2%)
3	delicious (15%/6%)	delicious (15%/8%)	location (11%/7%)	quality (10%/6%)
4	amazing (14%/7%)	definitely (17%/11%)	building (4%/1%)	excellent (25%/19%)
5	beautiful (11%/6%)	amazing (13%/8%)	premier (3%/1%)	good (35%/29%)
6	we (59%/50%)	beautiful (9%/5%)	town (5%/3%)	ale (3%/1%)
7	loved (6%/3%)	we (55%/47%)	inn (4%/2%)	girlfriend (1%/0%)
8	our (34%/26%)	boyfriend (1%/0%)	peas (1%/0%)	superb (6%/4%)
9	was (66%/58%)	fab (3%/1%)	quality (9%/6%)	class (3%/2%)
10	so (30%/23%)	so (28%/22%)	haddock (1%/0%)	well (23%/19%)
11	mum (2%/0%)	loved (6%/3%)	business (3%/1%)	pint (1%/0%)
12	had (41%/34%)	gorgeous (3%/1%)	centre (5%/3%)	in (57%/53%)
13	children (5%/2%)	love (7%/4%)	odd (1%/0%)	top (5%/4%)
14	recommend (22%/16%)	recommend (20%/16%)	nearby (3%/1%)	business (2%/1%)
15	fabulous (5%/2%)	fabulous (4%/2%)	importantly (1%/0%)	of (61%/57%)
16	love (6%/3%)	friend (11%/8%)	cheapest (1%/0%)	standard (4%/3%)
17	fab (3%/1%)	were (39%/34%)	budget (1%/0%)	great (36%/32%)
18	boyfriend (1%/0%)	yummy (1%/0%)	point (3%/1%)	establishment (2%/1%)
19	would (28%/22%)	would (25%/21%)	one (21%/17%)	real (4%/3%)
20	gorgeous (3%/1%)	our (28%/24%)	car (5%/3%)	that (25%/22%)

Table 3. Words that occurred disproportionately often in the 10-rated reviews of one gender in comparison to the other gender. The percentage of reviews containing the term for the majority and other gender is shown in brackets.

Rank	Hotels, female	Restaurants, female	Hotels, male	Restaurants, male
1	we (67%/49%)	husband (8%/1%)	wife (14%/1%)	wife (10%/1%)
2	even (27%/12%)	we (67%/56%)	avoid (9%/2%)	girlfriend (2%/0%)
3	were (65%/48%)	lovely (9%/4%)	required (4%/0%)	avoid (8%/5%)
4	daughter (5%/0%)	were (59%/50%)	hear (5%/1%)	poor (17%/13%)
5	most (9%/2%)	our (42%/33%)	sleeping (4%/0%)	beer (4%/2%)
6	member (12%/4%)	us (33%/26%)	four (8%/2%)	pint (2%/1%)
7	shocking (5%/0%)	boyfriend (2%/0%)	are (33%/20%)	is (47%/42%)
8	second (9%/2%)	didn't (20%/14%)	front (11%/4%)	quality (8%/6%)
9	returning (7%/1%)	was (84%/78%)	apart (4%/1%)	place (28%/25%)
10	wasn't (14%/6%)	rude (13%/9%)	road (4%/1%)	world (2%/1%)
11	on (64%/51%)	disappointed (11%/7%)	pretty (4%/1%)	best (7%/5%)
12	happened (4%/0%)	came (21%/15%)	stare (3%/0%)	simply (3%/2%)
13	decent (4%/0%)	so (44%/37%)	cleaning (3%/0%)	ale (1%/0%)
14	ruined (4%/0%)	couldn't (10%/7%)	piece (8%/2%)	worst (10%/8%)
15	had (65%/52%)	birthday (6%/4%)	order (21%/11%)	mate (1%/0%)
16	us (36%/24%)	had (60%/54%)	per (5%/1%)	you (34%/31%)
17	declined (4%/0%)	wasn't (12%/9%)	pint (5%/1%)	establishment (4%/3%)
18	sheet (4%/0%)	children (6%/4%)	future (5%/1%)	probably (4%/3%)
19	tried (11%/4%)	asked (24%/19%)	arrive (6%/2%)	adjacent (1%/0%)
20	dessert (7%/1%)	when (40%/34%)	milk (4%/1%)	low (2%/1%)

The key patterns found in Table 2 and 3 are summarised below.

- In hotel reviews with the highest ratings, women used more simple explicitly positive words. These included *lovely* (31% of female reviews, 17% of male reviews), *delicious* (15% of female reviews, 6% of male reviews), *amazing* (14% of female reviews, 7% of male reviews), *beautiful* (11% of female reviews, 6% of male reviews), *loved* (6% of female reviews, 3% of male reviews), *fabulous* (5% of female reviews, 2% of male reviews). There were no common explicitly positive words that were used more often by males. The closest examples were *quality* (9% of male reviews, 6% of female reviews) and *cheapest* (0.9% of male reviews, 0.1% of female reviews). Male reviews instead tended to use more common factual words than female reviews, such as *beer* (6% of male reviews, 2% of female reviews), *location* (11% of male reviews, 7% of female reviews), and *building* (4% of male reviews, 1% of female reviews).
- In hotel reviews with the lowest ratings, women used more simple explicitly negative words. These included *shocking* (5% of female reviews, 0% of male reviews), and *ruined* (4% of female reviews, 6% of male reviews). There were no common explicitly negative words that were used more often by males. The closest example was *avoid* (9% of male reviews, 2% of female reviews).
- In restaurant reviews with the highest ratings, women used more simple explicitly positive words. These included *lovely* (31% of female reviews, 16% of male reviews), *delicious* (15% of female reviews, 8% of male reviews), *amazing* (13% of female reviews, 8% of male reviews), *beautiful* (9% of female reviews, 5% of male reviews),

fab (3% of female reviews, 1% of male reviews), *loved* (6% of female reviews, 3% of male reviews), *gorgeous* (3% of female reviews, 1% of male reviews), *love* (7% of female reviews, 4% of male reviews), *recommend* (20% of female reviews, 16% of male reviews), *fabulous* (4% of female reviews, 2% of male reviews), *yummy* (1.3% of female reviews, 0.4% of male reviews). There were also several common explicitly positive words that were used more often by males. These included *excellent* (25% of male reviews, 19% of female reviews), *good* (35% of male reviews, 29% of female reviews), *superb* (6% of male reviews, 4% of female reviews), *top* (5% of male reviews, 4% of female reviews), *great* (36% of male reviews, 32% of female reviews).

- In restaurant reviews with the lowest ratings, women used simple explicitly negative words more. These included *rude* (13% of female reviews, 9% of male reviews), and *disappointed* (11% of female reviews, 7% of male reviews), although they also used one common positive word, *lovely* (9% of female reviews, 4% of male reviews). Males also used other explicitly negative words more. These included *poor* (17% of male reviews, 13% of female reviews) and *worst* (10% of male reviews, 8% of female reviews), as well as *avoid* (8% of male reviews, 5% of female reviews), and one explicitly positive term, *best* (7% of male reviews, 5% of female reviews).

From these results, there main reason for gender differences in accuracy seems to be the tendency for female reviewers to use common transparently positive words when awarding a high rating and, to a lesser extent, to use common transparently negative words when giving a low rating.

Discussion

The results are limited by using a single source for the data, the TripAdvisor website, and the dynamics of sentiment expression may differ in other contexts. There is likely to be an element of spam or paid reviews in TripAdvisor because of its commercial utility, which may have affected the results. Although the influence of spam should be minimised by restricting each dataset to at most one review for each reviewer, paid reviewers may have multiple TripAdvisor accounts. The star ratings of review sites are also problematic because they may be applied differently by reviewers (De Langhe, Fernbach, & Lichtenstein, 2016). The results apply only to lexical sentiment analysis and may not apply to sentiment analysis programs with different lexicons (e.g., Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). The answers to the research questions might be different for non-lexical sentiment analysis algorithms exploiting machine learning to learn indirect expressions of sentiment that associated with strong sentiment. It is also possible that the findings primarily relate to sentiment strength and would not be replicated for polarity detection applications. This latter seems reasonably likely due importance of strong expressions of sentiment to explain the gender differences found.

The findings do not align with a previous discovery that machine learning based sentiment polarity detection could be improved if terms that had sentiment associations that varied by gender were removed (Volkova, Wilson, & Yarowsky, 2013). Although a logical implication from this is that sentiment analysis could be more accurate when trained on single-gender datasets, insufficient evidence was found for this on the TripAdvisor datasets. Thus, the current paper does add to the evidence for the value of gender-specific sentiment analysis. It is possible that it is useful for the machine learning approach but not for the lexical approach.

In terms of the wider literature about emotions in the social web, the more frequent expression of positive sentiment by females found here confirms that females express positive sentiment more frequently in the social web (Thelwall, Wilkinson, & Uppal, 2010). The current paper deepens these findings with the discovery that this difference persists even when extended to balanced datasets in which males and females gave the same ratings to the products reviewed. Thus, the increased use of direct positive language by females is a communication *style* difference rather than a communication *substance* difference.

Conclusions

The results give evidence that lexical sentiment analysis algorithms have gender biases. They are likely to detect more strong positive and negative sentiment from females than from males. Thus, market researchers using social media monitoring software incorporating lexical sentiment analysis algorithms are likely to get misleading results when comparing male and female attitudes (e.g., the implicit assumption of equal accuracy for males and females in the separate sentiment-related scores for the two genders in Figure 9 of: Fuchs, Höpken, & Lexhagen, 2014). For example, a sentiment analysis of political tweets (e.g., Gul, Mahajan, Nisa, Shah, Jan, & Ahmad, 2016) might identify female concerns more readily than male concerns. Even when ignoring gender, the consumer sentiment information is likely to be misleading due to disproportionately highlighting traits that females are strongly positive or negative about. This issue should not be exaggerated because the gender differences are not large but in contexts when small differences are important, gender bias may lead to incorrect marketing decisions. Although the findings are specific to sentiment strength detection and the SentiStrength software, those using other software and approaches should test for similar problems before relying on their results. When this is not practical, users should be particularly sceptical of small sentiment differences concerning topics of more interest to women. Other applications of sentiment analysis (e.g., Skowron, Rank, Theunis, & Sienkiewicz, 2011; Weber, Ukkonen, & Gionis, 2012) should also consider gender limitations.

From a big data perspective, the conclusions add to the evidence that big data analyses in the social sciences, which often employ sentiment analysis, can give biased or misleading results. Although the potential for bias in big data analyses has previously been noted (Bollier, 2010; Boyd & Crawford, 2012; Hofacker, Malthouse, & Sultan, 2016; Tufekci, 2014), the source found here is more subtle than demographic differences in the people that express opinions online and is not due to analyst bias but is a hidden consequence of gender differences in communication styles. Gender bias may go beyond this if customer relations personnel are more easily able to identify dissatisfied female customers because their opinions are expressed more directly (e.g., see: Ma, Sun, & Kekre, 2015).

At the level of software, one way to circumvent the problem of gender bias in sentiment analysis would be to calibrate programs on a sample of accurate data and then apply a modifying factor to one of the genders to compensate (following the example of adjustment strategies for political polls: Wang, Rothschild, Goel, & Gelman, 2015). For example, if male top ratings were 10% lower than female ratings on an accurate sample then the top rating results for males on other data could be compensated by multiplying them by 10/9.

In terms of the wider issue of the role of gender in communication, the results suggest a small tendency for females to use more direct language to express the same sentiment as males. Whilst this is not surprising, it is a novel source of evidence that this difference operates at the level of communication rather than other factors, such as outlook on life.

References

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In LREC2010 (pp. 2200-2204).

Bagić Babac, M., & Podobnik, V. (2016). A sentiment analysis of who participates, how and why, at social media sport websites: How differently men and women write about football. Online Information Review, 40(6), 814-833.

Bollier, D. (2010). The promise and peril of big data. Washington, DC: Aspen Institute.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15(5), 662-679.

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating gender on Twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1301-1309). Association for Computational Linguistics.

De Langhe, B., Fernbach, P. M., & Lichtenstein, D. R. (2016). Navigating by the stars: Investigating the actual and perceived validity of online user ratings. Journal of Consumer Research, 42(6), 817-833.

Fuchs, M., Höpken, W., & Lexhagen, M. (2014). Big data analytics for knowledge generation in tourism destinations–A case from Sweden. Journal of Destination Marketing & Management, 3(4), 198-209.

Gronholdt, L., Martensen, A., & Kristensen, K. (2000). The relationship between customer satisfaction and loyalty: cross-industry differences. Total Quality Management, 11(4-6), 509-514.

Gul, S., Mahajan, I., Nisa, N. T., Shah, A., Jan, A. & Ahmad, S. (2016). Tweets speak louder than leaders and masses: An analysis of tweets about the Jammu and Kashmir elections 2014. Online Information Review, 40(7), 900-912.

Hofacker, C. F., Malthouse, E. C., & Sultan, F. (2016). Big data and consumer behavior: Imminent opportunities. Journal of Consumer Marketing, 33(2), 89-97.

Hofer-Shall, Z. (2010). The Forrester Wave: Listening Platforms, Q3 2010. Forrester Research. http://www.demainlaveille.fr/wp-content/uploads/2010/07/forrester.pdf

Homburg, C., Ehm, L., & Artz, M. (2015). Measuring and managing consumer sentiment in an online community environment. Journal of Marketing Research, 52(5), 629-641.

Kim, Y., Dwivedi, R., Zhang, J., & Jeong, S. R. (2016). Competitive intelligence in social media Twitter: iPhone 6 vs. Galaxy S5. Online Information Review, 40(1), 42-61.

Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4), 401-412.

Lee, H., & Suh, Y. (2016). Who creates value in a user innovation community? A case study of MyStarbucksIdea. com. Online Information Review, 40(2), 170-186.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.

Ma, L., Sun, B., & Kekre, S. (2015). The squeaky wheel gets the grease: An empirical analysis of customer voice and firm intervention on Twitter. Marketing Science, 34(5), 627-645.

Mihalcea, R., & Garimella, A. (2016). What men say, what women hear: Finding gender-specific meaning shades. IEEE Intelligent Systems, 31(4), 62-67.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1-135.

Pekar, V., & Ou, S. (2008). Discovery of subjective evaluations of product features in hotel reviews. Journal of Vacation Marketing, 14(2), 145-155.

Rangel, F., & Rosso, P. (2016). On the impact of emotions on author profiling. Information processing & management, 52(1), 73-92.

Schweidel, D. A., & Moe, W. W. (2014). Listening in on social media: A joint model of sentiment and venue format choice. Journal of Marketing Research, 51(4), 387-402.

Skowron, M., Rank, S., Theunis, M., & Sienkiewicz, J. (2011). The good, the bad and the neutral: affective profile in dialog system-user communication. In: International Conference on Affective Computing and Intelligent Interaction, Springer Berlin (pp. 337-346).

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307.

Thelwall, M. & Stuart, D. (2006). Web crawling ethics revisited: Cost, privacy and denial of service. Journal of the American Society for Information Science and Technology, 57(13), 1771-1779.

Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544–2558.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social Web, Journal of the American Society for Information Science and Technology, 63(1), 163-173.

Thelwall, M., Wilkinson, D., & Uppal, S. (2010). Data mining emotion in social network communication: Gender differences in MySpace. Journal of the Association for Information Science and Technology, 61(1), 190-199.

Thelwall, M. (2016). Book genre and author gender: Romance> Paranormal-Romance to Autobiography> Memoir. Journal of the Association for Information Science and Technology.

Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. Journal of Marketing Research, 51(4), 463-479.

Tufekci, Z. (2014). Big questions for social media big data: representativeness, validity and other methodological pitfalls. In Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM14). https://arxiv.org/abs/1403.7400

Turney, P. D. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 417-424). Association for Computational Linguistics.

Volkova, S., Wilson, T., & Yarowsky, D. (2013). Exploring demographic language variations to improve multilingual sentiment analysis in social media. In EMNLP (pp. 1815-1827).

Volkova, S., & Yarowsky, D. (2014). Improving gender prediction of social media users via weighted annotator rationales. In NIPS 2014 Workshop on Personalization. http://www.cs.jhu.edu/~svitlana/papers/VY-NIPSPersonalization14.pdf

Volkova, S., & Yoram, B. (2015). On predicting sociodemographic traits and emotions from communications in social networks and their implications to online self-disclosure.

Cyberpsychology, Behavior, and Social Networking, 18(12), 726-736. doi:10.1089/cyber.2014.0609.

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. International Journal of Forecasting, 31(3), 980-991.

Weber, I, Ukkonen, A., & Gionis, A. (2012). Answers, not links: extracting tips from Yahoo! answers to address how-to web queries, Proceedings of the fifth ACM international conference on Web search and data mining (WSDM12). New York, NY: ACM Press (pp. 613-622).

Appendix

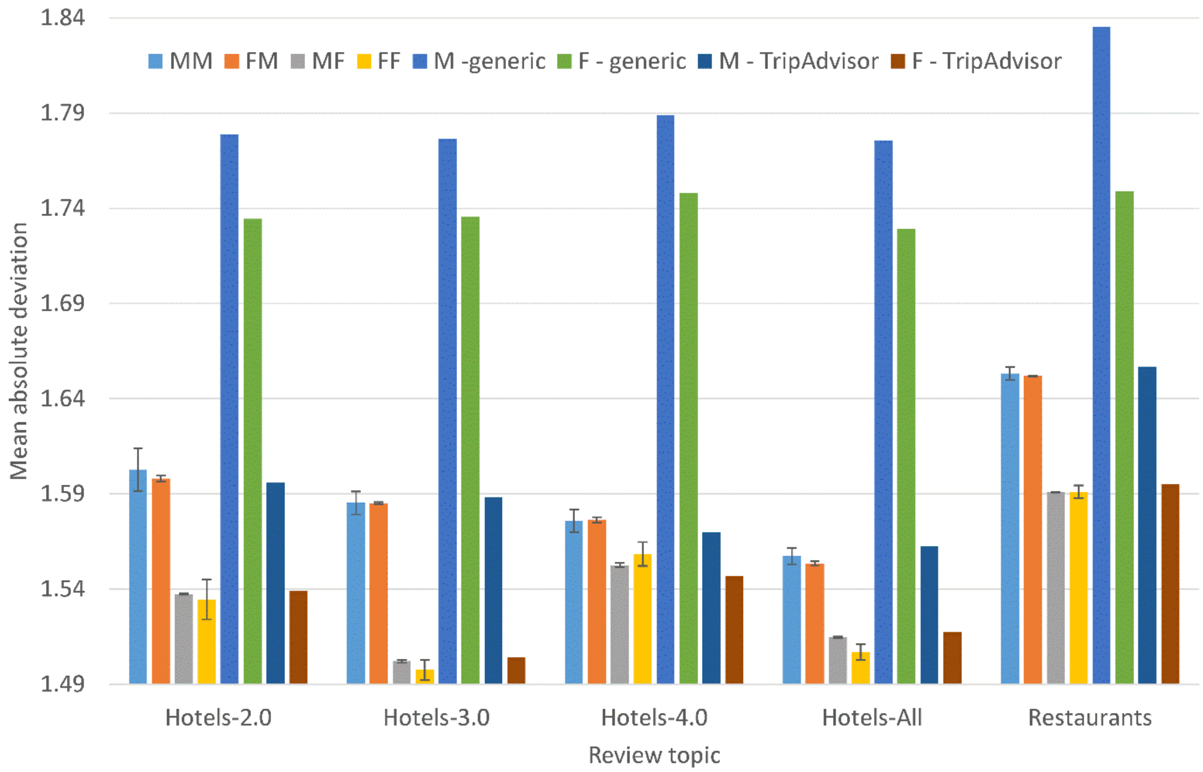

Figure 4. An expanded version of Figure 1, including the SentiStrength scores before overall optimisation on TripAdvisor data.