Analysis of Airline Sentiment Data: Does the social media image reflect real performance?

Haoran Zheng

Tianjin Yinghua International School Tianjin, China

Abstract.

With the development of social media, increasingly more people express their feelings on social media (such as Twitter), which are a useful source of information e.g. for the airline companies which want to find out what causes negative sentiment about them and try to improve the communicated weaknesses. This paper aims to find out the variation in sentiment towards different airlines and check whether the Twitter perception of airlines reflects their "real" performances. The result suggests that Delta and Southwest airlines have a relatively good reputation, while US Airways has a bad reputation. By examining the reasons for negative sentiment, we find that Twitter ranking reflects the "real" ranking based on the Department of Transportation data closely for mishandled luggage and canceled flights, but considerably less closely for delayed flights. This demonstrates that Twitter can provide a good reflection of reality, but this is not always the case.

Keywords: Airline sentiment, Social media, Twitter

3rd International Conference On Research In

MANAGEMENT AND ECONOMICS

1. Introduction

Airlines pay attention to how they are presented in social media. That involves user-generated content, such as tweets by the customers. Such information can be used for example to improve the quality of the service. However, the sentiment towards a particular airline may not always be representative of its real performance. One of the main reasons for a possible disparity is the randomness component involved. For example, despite a large number of tweets that mention the mishandling of luggage, it may be the case that the airline ranks highly in that particular respect in the official statistic. In other words, not always is the prevalence and severity of the problem reflected in the relatively small sample that ends up published on Twitter.

Knowledge of their perception of social media is important to airlines for several reasons. First, it allows them to understand the source of negative sentiments so that they can think about improvements in their service. Second, the airline may tailor its marketing strategy to address the perceived weaknesses of their service.

This paper presents a study that uses social media to analyse the data of airline sentiment. The dataset consists of tweets that mention one of the six major US airlines, which are classified into positive, neutral, or negative. If the tweet falls into the last category, the reason for customer dissatisfaction is also extracted.

The goal of the study was twofold. First, we examine the differences between airline sentiment and the reasons for them. Seconds, we check whether the Twitter perception of an airline is reflected in the "real" statistics on the airline's performance.

We use a dataset that processed 14,460 publicly shared tweets, which has categorized the sentiment for each tweet. If the sentiment was negative, the dataset has also extracted the reason for it. We calculate the prevalence of positive- and negative-sentiment tweets for each airline and check if there are meaningful differences between them. Moreover, for the subset of negative tweets, we perform a similar exercise for the set of reasons that led to consumer dissatisfaction.

For the second goal of the paper, we introduce another dataset, which is the US Department of Transportation statistics. This shows the real punctuality as well as the quality of luggage handling for each airline for the same period as the Twitter dataset. This allows us to compare whether the prevalence of specific reasons for customer dissatisfaction is reflected in the prevalence of negative tweets. In other words, we examine to what extent the image of the airline that the customer obtains from Twitter is representative of reality.

The plan for the paper is as follows. First, we review the related literature. Then, we introduce and describe the data used. Next, we analyse and discuss the hypotheses stated above. Finally, we conclude by assessing the reliability and validity of the information on Twitter.

2. Literature review

We begin by a brief characterisation of social media. (O'Leary, 2011) says that social media is the availability of free, accessible, and easy-to-use online tools, that enable interactive exchanges between individuals and groups. Also, (Deshpande & Sarkar, 2010) claim that social media is characterized by honesty, authenticity and originality, and consumer views are one of the most popular types of information online. (Garg et al., 2011) demonstrate that Twitter is among the world's top ten social media channels in terms of registered users and daily visits. Finally, (Savage, 2011) and (Thelwall et al., 2011) note that Twitter is one of the most popular channels for sharing information and receiving real-time updates.

(Misopoulos et al., 2014) claim that sentiment analysis allows researchers to identify subjects of opinions, classify them according to the viewpoints and determine their emotional connotation, thus enabling researchers to determine what aspects of customer experiences elicit particular responses and feelings. They analyse sentiment towards the airlines using the following approach. Firstly, collected tweets are classified into three pools of data, by determining whether their content included one of the following keywords: "good", "fail", or "lounge". The purpose of the last keyword is to create a category of tweets that capture the sentiments of the customers using frequent flier or business class programs offered by two full-frills airlines. In the second stage, called "term weighting", the frequency of occurrence is calculated for each term (dimension). In the third stage, the query criteria are matched with the terms (keywords) within a tweet, thus indicating similarities.

By analysing Twitter feeds they are able to identify a common pattern among customer complaints, discover the source of negative sentiments, and review specific problems encountered by customers in this area of service delivery. Similarly, the approach also successfully uncovers service features and practices that caused positive sentiments among customers. Customers of airlines are eager to share both positive and negative aspects of their service experiences, and are often directly addressing companies with requests to address their problems. Their paper contains 5.34% positive tweets, 4.44% negative tweets, and 90.2% neutral tweets, which shows an overwhelming majority of neutral comments and more positive comments than negative comments. This contrasts with the descriptive statistics for my dataset, where there are 62.69% negative tweets, 21.17% neutral tweets and 16.14% positive tweets. The differences may be explained by a different classification method used; in particular, the algorithm in (Misopoulos et al., 2014) always errs on the side of caution, assigning tweet to the neutral category in case of any uncertainly.

(Wenjing Duan et al., 2013) and (Parasuraman et al., 1985) conduct a similar study concerning hotels. They state that the service quality evaluated by consumers comprises the same five dimensions: tangibles, reliability, responsiveness, assurance, and empathy. The authors also use a sentiment analysis technique and find that tangibles seem to have the

strongest impact on user review ratings. This finding suggests that online users' review behaviour in the hotel industry is significantly driven by their experience in these five major areas.

(Hodeghatta, 2013) analyses movie sentiment on Twitter. First, downloads Hollywood movie-related tweets from the Twitter automatically. The dataset includes data from nine countries over a forty-five period. Next, a sentiment analyser tool is developed using Python and natural language tool kit libraries by trying four different supervised machine learning algorithms. He also classifies these comments into positive and negative sentiment. Moreover, he analyses various consumer views on any specific aspect of the movie, which are useful foundations for companies' strategies. This paper finds that different regions express different sentiments depending on the nature of the movie and how the movies impact cultural sentiments.

(Spark & Browning, 2014) suggest that people often turn to the Internet to complain if the problem is not well addressed by the supervisor. (Kowalski & Erickson, 1997) identify four functions of complaining behaviour: catharsis (venting); self-presentation (impression management); social comparison (testing one's perception against others); and call for accounts (redress tactic). Also, (Price et al., 1995) find a significant relationship between altruism and what they termed market helping behaviour. It seems that the severity of the situation for the complainant did not necessarily influence whether the complaint was posted or not. For instance, uncomfortableness, concern for their health and safety, and financial loss have all featured heavily in the results despite considerably different severity.

3.Data

The data in our paper cover all tweets which reference a US-based airline during the approximately one-week period of 16th-24th February 2015. The data comes from Kaggle, which had collected it from Twitter through its API. In total, 14,460 tweets have been gathered.
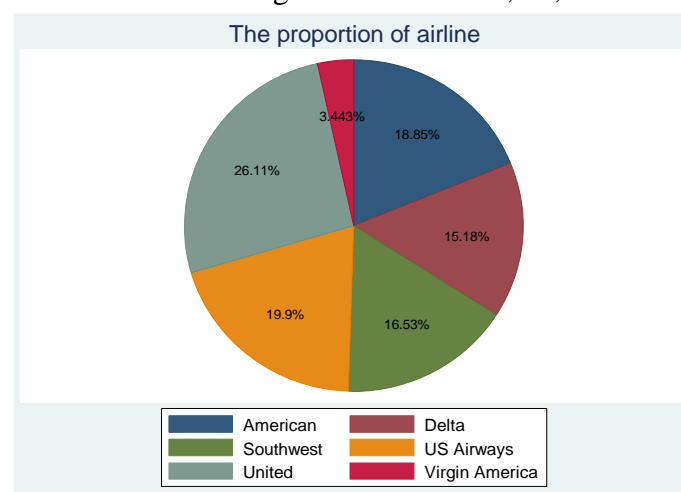

Figure 1: The proportion of tweets relating to each airline

Each tweet addresses one of the six airlines covered by our paper: United, US Airways, American, Southwest, Delta, and Virgin America. As shown in Figure 1, United had the largest presence at 26.1%. The other airlines constitute 15-20% of the tweets each, except Virgin America which constitutes 3.4%.
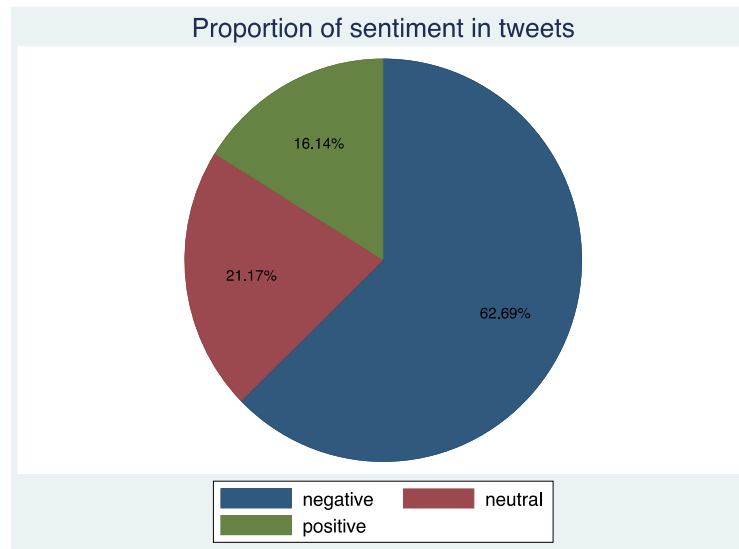

Figure 2: The classification of tweets into positive, neutral and negative

The tweets have been classified as positive, negative, or neutral comments based on the textual analysis. If the classification algorithm was uncertain about whether the tweet was positive or negative, it would have been classified as neutral. Among all tweets, negative comments have the highest percentage (62.7%), followed by neutral comments (21.2%) and positive comments (16.1%), as shown in Figure 2. This is not surprising as studies have shown that people are more likely to post negative reviews than positive or (deliberately) neutral ones.

In addition to classifying the sentiment of the tweet, for negative tweets the algorithm has also attempted to classify the stated reason for dissatisfaction. Nine possibilities have been identified, and an additional category called "Can't tell" was created for tweets for which classification was not possible. The reasons are listed below:

- **Bad flight:** everything that annoying people during the flight like the damaged ventilation or no toilet paper in the loo;
- **Cancelled flight;**
- **Customer service issue:** maybe somebody is rude; cannot change something about the flight;
- **Damaged luggage;**
- **Flight attendant complaints:** not good service of attendants; they are rude;
- **Flight booking problems:** the website is slow/not working or troublesome;

- **Late flight;**
- **Lost luggage;**
- **Long lines:** too many people waiting in line.

Table 1: Frequency of reasons for negative comments

```
. tabulate negativereason
```

negativereason	Freq.	Percent	Cum.
Bad Flight	580	6.32	6.32
Can't Tell	1,190	12.97	19.29
Cancelled Flight	847	9.23	28.51
Customer Service Issue	2,910	31.71	60.22
Damaged Luggage	74	0.81	61.03
Flight Attendant Complaints	481	5.24	66.27
Flight Booking Problems	529	5.76	72.03
Late Flight	1,665	18.14	90.17
Lost Luggage	724	7.89	98.06
longlines	178	1.94	100.00
Total	9,178	100.00	

Table 1 shows some negative reasons that customers have posted on Twitter. The most common complaint that the airline customers have had was customer service issues (32%), followed by late flight (18%). In 13% of cases, it was impossible to determine the source of dissatisfaction. The remaining reasons were less popular with the smallest number of tweets citing long lines and damaged luggage. As mentioned above, the prevalence of comments may have been affected by the severity of the incidents faced by the customers.

4.analysis

We begin the analysis by examining the differences between Twitter sentiment for each airline. To do this, we create a two-way table that splits that tweets based on the airline and the recognized sentiment (positive, neutral, or negative).

Table 2: Two-way table of sentiment by airline

airline	airline_sentiment			Total
	negative	neutral	positive	
American	1,960	463	336	2,759
	71.04	16.78	12.18	100.00
Delta	955	723	544	2,222
	42.98	32.54	24.48	100.00
Southwest	1,186	664	570	2,420
	49.01	27.44	23.55	100.00
US Airways	2,263	381	269	2,913
	77.69	13.08	9.23	100.00
United	2,633	697	492	3,822
	68.89	18.24	12.87	100.00
Virgin America	181	171	152	504
	35.91	33.93	30.16	100.00
Total	9,178	3,099	2,363	14,640
	62.69	21.17	16.14	100.00

Table 2 shows the split between the negative, neutral and positive comments for each airline. US Airways had the highest percentage of negative comments at 77.7% while Virgin America had the lowest percentage of 36%. Similarly, US Airways had the lowest percentage of positive comments at 9.2%, while Virgin America had the highest percentage of 30.1%. It indicates that Virgin America has the best reputation, while US Airways the worst reputation among the airlines analysed in the paper.

To confirm that the differences in the number of negative tweets are meaningful, we compare the observed number of negative tweets for every airline with the expected number of negative tweets. The latter is calculated as the overall percentage of negative tweets in the whole sample multiplied by the total number of tweets about the particular airline. We conduct a chi-squared test of whether the differences are statistically significant. As can be seen in Table 3, we firmly reject the hypothesis that all airlines have the same proportion of negative comments. We also repeat the same exercise for positive tweets, and we reach identical conclusions.

Table 3: Chi-squared test for the equal proportion of negative tweets among all airlines

```
         Pearson chi2(5) = 425.3041   Pr =  0.000
likelihood-ratio chi2(5) = 455.2940   Pr =  0.000
```

observed	expected	obs – exp	Pearson
1960	1729.000	231.000	5.555
955	1392.000	–437.000	–11.713
1186	1517.000	–331.000	–8.498
2263	1826.000	437.000	10.227
2633	2396.000	237.000	4.842
181	315.000	–134.000	–7.550

Table 4: Two-way table of negative reason by airline

negativereason	American	Delta	Southwest	US Airw..	United	Virgin ..	Total
Bad Flight	87	64	90	104	216	19	580
	15.00	11.03	15.52	17.93	37.24	3.28	100.00
	4.44	6.70	7.59	4.60	8.20	10.50	6.32
Can't Tell	198	186	159	246	379	22	1,190
	16.64	15.63	13.36	20.67	31.85	1.85	100.00
	10.10	19.48	13.41	10.87	14.39	12.15	12.97
Cancelled Flight	246	51	162	189	181	18	847
	29.04	6.02	19.13	22.31	21.37	2.13	100.00
	12.55	5.34	13.66	8.35	6.87	9.94	9.23
Customer Service Is..	768	199	391	811	681	60	2,910
	26.39	6.84	13.44	27.87	23.40	2.06	100.00
	39.18	20.84	32.97	35.84	25.86	33.15	31.71
Damaged Luggage	12	11	14	11	22	4	74
	16.22	14.86	18.92	14.86	29.73	5.41	100.00
	0.61	1.15	1.18	0.49	0.84	2.21	0.81

Flight Attendant Co..	87	60	38	123	168	5	481
	18.09	12.47	7.90	25.57	34.93	1.04	100.00
	4.44	6.28	3.20	5.44	6.38	2.76	5.24
Flight Booking Prob..	130	44	61	122	144	28	529
	24.57	8.32	11.53	23.06	27.22	5.29	100.00
	6.63	4.61	5.14	5.39	5.47	15.47	5.76
Late Flight	249	269	152	453	525	17	1,665
	14.95	16.16	9.13	27.21	31.53	1.02	100.00
	12.70	28.17	12.82	20.02	19.94	9.39	18.14
Lost Luggage	149	57	90	154	269	5	724
	20.58	7.87	12.43	21.27	37.15	0.69	100.00
	7.60	5.97	7.59	6.81	10.22	2.76	7.89
longlines	34	14	29	50	48	3	178
	19.10	7.87	16.29	28.09	26.97	1.69	100.00
	1.73	1.47	2.45	2.21	1.82	1.66	1.94
Total	1,960	955	1,186	2,263	2,633	181	9,178
	21.36	10.41	12.92	24.66	28.69	1.97	100.00
	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table 4 shows that customer service presented the biggest issues for customers of all the airlines except Delta. Their proportion of all complaints ranged from 39.18% (for American) to 20.84% (for Delta). For Delta, the biggest issue mentioned was late flights (28.1%), which was the second or third most important issue faced by customers of other airlines. The remaining most important issue was cancelled flights, which were particularly important for American, Southwest, and Virgin. Other issues (including lost luggage) were less prevalent amongst the negative tweets.

Table 4 also shows that United Airlines had the largest share of all these negative reasons except customer service issues, long lines, and cancelled flights. Their proportion of all complaints ranged from 37.24% (for bad flight) to 21.37% (for cancelled flight). For customer service issues and long lines, US Airways had the biggest share (27.87% and 28.09% respectively), and also had the second largest share of all negative reasons. For cancelled flights, American had the biggest share (29.04%), which has the third largest share of all negative reasons. Other airlines all have a smaller share of all negative reasons.

We will now compare the image of reality that users may get from Twitter to the official statistics about the airlines' performance. The differences may reflect the issue with Twitter: people tend to report negative experiences more often than positive or neutral experiences. This is most likely due to the impulse nature of Twitter usage, i.e. it is easy and quick to convey one's emotions of anger or disappointment. This is particularly true for airlines as opposed to e.g. restaurants because it is difficult for the airline to exceed expectations while even one problem can lead to a negative reaction. Moreover, the proportion of negative tweets citing a particular problem reflects not only how often this problem occurs but also how seriously it affects the customers. That is why even if for example long queues at airports were the most prevalent problem, we may not see this from the Twitter data, as passengers are used to them, so they are not treated as seriously as e.g. a cancelled flight.

Table 5: Prevalence of selected negative reasons (delayed flight, cancelled flight, mishandled luggage) in Twitter data for each airline

Airline	Delayed (%)	Cancelled (%)	Mishandled Luggage (number per 1000 passengers)
American Airlines	12.7	12.55	58.4
Delta	28.17	5.34	30.6
Southwest	12.82	13.66	43.0
US Airways	20.02	8.35	56.6
United	19.94	6.87	76.1
Virgin America	9.39	9.94	17.9

Table 5 shows the negative reasons (delayed flight, cancelled. flight, mishandled luggage) for 6 airlines from Twitter. For the delayed flight, Delta had the highest percentage at 28.2%. It was followed by US Airways (20.0%) and United (20.0%). Virgin America had the lowest percentage at 9.4%. For the cancelled flight, Southwest has the highest percentage at 13.7%, followed by American Airlines (12.6%). Delta had the lowest percentage at 5.3%. For mishandled luggage, United had the number of 76.1 per 1000 passengers. Followed by American Airlines (58.4). Virgin America had the smallest number of 17.9 per 1000 passengers.

Table 6: Prevalence of selected negative reasons (delayed flight, cancelled flight, mishandled luggage) in US Department of Transportation data for each airline

Airline	On-time (%)	Delayed (%)	Cancelled (%)	Mishandled Luggage (number per 1000 passengers)
American Airlines	73.44	19.89	6.41	5.22
Delta	77.99	19.03	2.79	3.1
Southwest	77.54	18.44	3.83	3.36
US Airways	72.75	21.8	5.22	4.65
United	73.63	23.43	2.78	3.86
Virgin America	76.13	19.87	3.67	1.04

Table 6 shows negative reasons (delayed flight, cancelled flight, mishandled luggage) for 6 airlines from the US Department of Transportation in the same period when the Twitter data was gathered. For the delayed flight, United had the highest percentage at 23.4%, followed by

US Airways (21.8%). Southwest had the lowest percentage of 18.4%. For the canceled flight, American Airlines had the highest percentage at 6.41%, followed by Us Airways of 5.22%. Delta and United had low percentages of 2.79% and 2.78% respectively. For mishandled luggage, American Airlines had the biggest number of 5.22 per 1000 passengers, followed by US Airways (3.86). Virgin America had the smallest number of 1.04.

Table 7: Comparison of Twitter and the US Department of Transport data. Ranked from highest to lowest. Same colour maintained throughout the table for each airline.

	Delayed flights (%)			Cancelled flights (%)			Mishandled Luggage (per 1000 passengers)	
Ranking	**Twitter**	**US DoT**		**Twitter**	**US DoT**		**Twitter**	**US DoT**
1	Delta 28.17	United 23.43		Southwest 13.66	American Airline 6.41		United 76.1	American Airline 5.22
2	US Airways 20.02	US Airways 21.8		American Airline 12.55	US Airways 5.22		American Airline 58.4	US Airways 4.65
3	United 19.94	American Airline 19.89		Virgin 9.94	Southwest 3.83		US Airways 56.6	United 3.86
4	Southwest 12.82	Virgin 19.87		US Airways 8.35	Virgin 3.67		Southwest 43.0	Southwest 3.36
5	American Airline 12.7	Delta 19.03		United 6.87	Delta 2.79		Delta 30.6	Delta 3.1
6	Virgin 9.39	Southwest 18.44		Delta 5.34	United 2.78		Virgin 17.9	Virgin 1.04

These two datasets are merged in Table 7. Note that the numbers may not be directly compared between Twitter and US Department of Transportation, as they are expressed in different units. However, the ranking of airlines can be compared we can see some differences between Twitter and the US Department of Transport datasets.

For delayed flights, US Airways has the same rank in both Twitter and US Department of Transportation datasets (US DoT). United, Southwest, American Airline and Virgin have two rank differences between Twitter dataset and US DoT dataset, for instance, United ranks third in Twitter while first in US DoT. Moreover, Delta shows considerably different performance: it ranks first in Twitter while fifth in US DoT.

For cancelled flights, American Airline, Virgin, United and Delta all have one rank difference between two datasets, while Southwest and US Airways have two rank differences.

For mishandled luggage, Southwest, Delta and Virgin match well, which all have the same rank between two datasets. American Airline and US Airways both have one rank difference between two datasets, while United has two rank differences.

5.Conclusion

Results from our study suggest some meaningful differences in sentiment towards airlines between major six players in the US market. First, Virgin America has the highest percentage of positive comments and the lowest percentage of negative comments; thus, Virgin America seems to have the best reputation. However, the number of tweets about Virgin America is small (only 3.4% of all tweets), so it may not reflect the real performances exactly. Other than that, Delta and Southwest have relatively high percentages of positive comments and low percentages of negative comments. These two airlines have a relatively good reputation. In contrast, American, United, and US Airways have a low percentage of positive comments and a high percentage of negative comments, so these three airlines have a relatively bad reputation. Among these three airlines, US Airways has the lowest percentage of positive comments and the highest percentage of negative comments. We find the differences between in proportions of negative and positive tweets to be statistically significant between airlines.

By comparing Twitter and the US Department of Transport datasets, we conclude that Twitter is somewhat indicative of the real performance of the airlines, but the data also often diverge. For mishandled luggage, the rankings given by these two datasets are almost coincident. For cancelled flights, they are relatively coincident but sometimes there would be minor deviations. For delayed flights, the two rankings do not match well. This phenomenon may be caused by different definitions of a "delayed" flight between the public and the official data. In particular, customers may think the flight is late just because it is delayed for a few minutes, while the official definition stipulates that only if the actual arrival time is 15 minutes more than planned time, the flight is late. Therefore, it is natural for these two rankings to diverge, as they measure different types of events.

The survey of literature has shown that people are more likely to post negative comments on social media platforms. If the journey goes smoothly, passengers take it for granted and do not bother to post positive comments. But whenever there are some problems, passengers will be relatively angry and send negative comments. As a result, negative comments tend to be much more common than positive ones. Also, passengers put different emphasis on the reasons for each negative comment. People tend to take more severe causes seriously. The passenger may have encountered many problems, but he may have only posted about what he considers the most serious ones, such as flight cancellations. This mechanism is another reason why we may

observe a divergence between social media image and real statistics. Such a phenomenon is not specific to the airline industry, but other industries are likely to experience it as well.

In fact, an important limitation of this paper is that we only focus on the airline industry. Also, the dataset is a little old (it was compiled in 2015) and the specific time interval of this dataset is only around a week. Therefore, we may only see part of the picture; for example, a particular airline might have experienced a particularly good or bad streak during the short time when data was collected. If this were the case, our results would not be representative of its long-term performance. This raises questions about how representative our sample is of the overall airline performance. The problem could be mitigated by collecting more data over a longer time period. Moreover, the parsing algorithms used in this paper are relatively simple and we rely on the pre-processing of the data. Consequently, we cannot fully assess the quality of sentiment classification (or the classification of the negative reasons).

For future research, a more sophisticated and advanced sentiment analysis technique could be developed and applied in future papers. Moreover, further research should be conducted in other industries where social media posts can be used to analyse sentiment. Finally, future papers could compare different datasets of sentiments coming from different social media – not only Twitter - so that we can analyse the relationship between them, as well as whether the combined image they give lies closer to reality.

References

Deshpande, M., & Sarkar, A. (2010). BI and sentiment analysis. *Business Intelligence Journal*, *15*(2), 41-49.

Garg, R., Smith, M. D., & Telang, R. (2011). Measuring information diffusion in an online community. *Journal of Management Information Systems*, *28*(2), 11-38.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, *62*(2), 406-418.

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1985). A conceptual model of service quality and its implications for future research. *Journal of Marketing*, *49*(4), 41-50.

Sparks, B. A., & Browning, V. (2010). Complaining in cyberspace: The motives and forms of hotel guests' complaints online. *Journal of Hospitality Marketing & Management*, *19*(7), 797-818.

Price, L. L., Feick, L. F., & Guskey, A. (1995). Everyday market helping behavior. *Journal of Public Policy & Marketing*, *14*(2), 255-266.

Duan, W., Cao, Q., Yu, Y., & Levy, S. (2013, January). Mining online user-generated content: using sentiment analysis technique to study hotel service quality. In *2013 46th Hawaii International Conference on System Sciences* (pp. 3119-3128). IEEE.

Hodeghatta, U. R. (2013, August). Sentiment analysis of Hollywood movies on Twitter. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)* (pp. 1401-1404). IEEE.

Guercini, S., Misopoulos, F., Mitic, M., Kapoulas, A., & Karapiperis, C. (2014). Uncovering customer service experiences with Twitter: the case of the airline industry. *Management Decision*.

O'leary, D. E. (2011). The use of social media in the supply chain: Survey and extensions. *Intelligent Systems in Accounting, Finance, and Management*, *18*(2-3), 121-144.

Savage, N. (2010). New search challenges and opportunities. *Communications of the ACM*, *53*(1), 27-28.

Kowalski, R. M., & Erickson, J. R. (1997). What's all the fuss about. *Aversive Interpersonal Behaviors*, *91*.