Reliability of Sentiment Mining Tools: A comparison of Semantria and Social Mention

Author: Linus Philip Lawrence
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
l.p.lawrence@student.utwente.nl

Social Media platforms have become quintessential for user-generated content and consumer opinions. As a result, vast amounts of commercially and freely available sentiment mining tools have emerged, however it remains unclear how reliable these tools are. In this paper, I evaluate the reliability as well as the features of the sentiment mining tools Semantria and Social Mention. This study is a sentiment analysis of 12 different car models that were applied to both sentiment mining tools. In addition to presenting a comparison of outputs obtained from two different sentiment mining tools, an analysis was conducted that compares the sentiment and passion outputs for three social media platforms. The results show significant differences in outputs for sentiment mining tools. Furthermore, statistically signifcant as well as observational differences were also apparent for the outputs from three social media platforms. The results have theoretical as well practical implications for different groups, including academics and practitioners using sentiment mining tools as input for future research and decision-making.

Supervisors: Dr. A.B.J.M. (Fons) Wijnhoven

Keywords
Sentiment analysis, Social Media, Twitter, Sentiment mining tools, Social Mention, Reliability

1. INTRODUCTION

Social media platforms such as Facebook, MySpace, Twitter and many other platforms, have facilitated an explosion of online user-generated content. The content from these social media platforms is publicly available and provides opportunities for tapping into a user's social network, preferences and opinions about products, services, or people (Tang & Liu, 2010). This availability of large-scale social data has turned the web into a potential source of information for social science research (Thelwall, Wouters & Fry, 2008). However, social data is also being exploited for commercial purposes in order to extract consumer opinions about products or brands (Thelwall, Buckley & Paltoglou, 2011). With this data, one can use social media to gauge customer preferences of what they like or dislike (Ostrowski, 2010).

One of the main factors of understanding and interpreting social media data is analysing customer sentiment. This term can be defined as "the consensus of feelings that consumers have about a product" (Ostrowski, 2010, p.1). The method of extracting customer sentiment falls in the field of sentiment analysis, also known as opinion mining or sentiment mining. It is the task of identifying whether any opinion expressed, in the form of a document, sentence or phrase, is positive or negative (Pang & Lee, 2008) and is predominantly conducted with the help of a sentiment mining tool. This fact that any kind of sentiment mining is the result of a tool and platform brings forth the issue of reliability and validity problems. The issue of reliability concerns itself whether the results from two sentiment mining tools are consistent throughout the period they are tested. This is the so-called internal consistency reliability. Issues of validity concern themselves with the degree to which a test really measures what it claims to assess. In other words, what have I seen of reality and is the data a good representation of reality? This is the so-called construct validity (Zachariadis, Scott & Barrett, 2013).

The growing interest in sentiment mining has led to numerous commercially and freely available sentiment mining tools having been developed. One of these tools is Social Mention[1] , which is a widely used social media search and analysis tool (see Section 4.2). The other is Semantria[2], a sentiment analysis tool that can be applied directly to any kind of documents, sentences or phrases. As the nature of these sentiment mining tools is very different, my research examines the different outputs that are generated by both tools. Therefore, my first research question examines:

How do sentiment ratios vary for different sentiment mining tools?

Furthermore, Social Mention automatically aggregates sentiment indicators for different social media platforms. This leads me to my second research question:

How do Social Mention's metrics vary for different social media platforms?

To answer these research questions, I collected social media data over a period of one month. In order to compare the results of sentiment mining tools and different social media platforms, I collected data for 12 different car models. Each car model served as a control variable for three different social media platforms. This allowed me to compare the results for different sentiment mining tools, social media platforms and also provided an insight into each car model's social media presence.

[1] www.socialmention.com/

[2] https://semantria.com/

Towards my goal, there a number of considerations to be made about social media and its viability for sentiment analysis. Firstly, it is crucial to recognise whether customers speak and share their 'real' opinions on social media. Is this not the case, then there is no reason and consequently no value in analysing customer sentiments from social media. However, the so-called "online disinhibition effect" (Suler, 2004; Joinson, 2007), makes people more likely to share their own opinion in an online setting than they would in direct communication with other people. This phenomenon makes me believe, if people post their opinions online, this publicly accessible data can be used to identify and monitor social media patterns. Secondly, any brand or product could have been chosen as the sample in this paper. However, I consider an investigation within the automotive sector to be of great interest as it is, (1) fundamentally relevant for the broad masses of young people to senior citizens, and (2) there exists a high amount of social media marketing aimed at this sector (Ostrowski, 2010).

With this paper, I do not merely aim to answer the research questions and confirm or reject the stated hypotheses. A further contribution is made by indicating the problems and limitations when using social media sentiments and social media analysis tools. By gathering a dataset of roughly 12,000 tweets for 12 different car models, I calculated the significant difference of outputs for two different sentiment mining tools and three social media platforms. In addition, I aim to explain biases and causes of the results by means of analysing each car models social media campaigns.

The rest of this paper is organised as follows. Section 2 discusses the existing body of knowledge on sentiment analysis (opinion mining) and the application of sentiment analyses for real-world situations. Section 3 describes a model for comparing two sentiment mining tools. I describe both sentiment mining tools, the data collection of social media data and the methodology in Sections 4 and Section 5. The results of the comparison of Semantria and Social Mention are presented in Section 6. Lastly, I critically reflect upon my research with the discussion in Section 7 and conclude about my work in Section 8.

2. RELATED WORK

2.1 Sentiment Analysis

Sentiment analysis, also known as opinion mining, is the task of identifying opinions expressed in texts and whether the expressions indicate positive or negative opinions toward a subject or topic (Nasukawa & Yi, 2003). According to Nasukawa and Yi, sentiment analysis involves the identification of sentiment expressions, polarity and strength of the expressions, and the relationship to the subject or topic.

Over the last decade, there has been a growing interest in the area of Natural Language Processing, the main aspect of sentiment analysis. Research ranges from document-level classification (Pang and Lee,2008) to studying the polarity of words and phrases (Wilson, Wiebe and Hoffmann, 2005; Esuli and Sebastiani, 2006).

Given the nature of tweets, i.e. limited to 140 characters, the sentiment of these messages can be classified of being similar to sentence-level sentiment analysis as described by Yu and Hatzivassiloglou (2003), and Kim and Hovy (2004). However, Kouloumpis, Wilson and Moore (2011) state that "the informal and specialized language used in tweets, as well as the very nature of the micro-blogging domain make Twitter sentiment analysis a very different task" (p. 1). A common question that

arises is, how well do the features and techniques used on more structured (well-formed) data transfer to the micro-blogging domain? The inherent characteristics of the micro-blogging domain, which includes Twitter, provide much more unstructured (ill-formed) data, compared to structured data that can be extracted from documents and phrases. The use of informal language (internet slang), emoticons, hashtags, @ and abbreviations such as FTW (for the win), are the main causes of unstructured data. This has led to researchers taking different approaches when using micro-blogs, such as Twitter, for sentiment analysis purposes. A number of papers have used emoticons for defining training data in order to recognise positive and negative sentiments in tweets (Read, 2005; Go et al., 2009; Pak & Paroubek, 2010). The authors of these publications used the Naive Bayes classifier to form training sets for sentiment classification. Davidov, Tsur and Rappoport (2010) used emoticons and hashtags for creating a training set for sentiment/ non-sentiment classification.

A different approach of sentence-level sentiment analysis of micro-blogging data is to concentrate on words and phrases being used. Asur and Huberman (2010), Fiaidhi et al (2012), Aston, Liddle and Hu (2014) use sentiment analysis tools to conduct sentence-level analyses. Examples of such tools are Python NLTK (Natural Language Toolkit)[3], R (text mining module)[4], RapidMiner[5] which are all freeware, but require a degree of coding. However, there are also fee-based sentiment analysis tools, such as Semantria, Lingpipe[6] and LIWC2007 (Linguistic Inquiry and Word Count)[7] that are mostly automated and only require the datasets to be uploaded. All of these tools predominantly use dictionary-based and machine learning techniques, i.e. using data found in lexicographical resources to assign sentiments to a large number of words. Together with unknown algorithms, these tools determine the polarity of a given document-whether the expressed opinion is positive, negative or neutral. However, most sentiment analysis tools offer an advanced sentiment classification that goes 'beyond polarity'. For instance, some tools analyse emotional states such as sad, happy or angry, whereas others determine the sentiment of a sentence with specific sentiment scores. The abundance of sentiment analysis tools allows researchers as well as business professionals or freelancers to conduct sentiment mining for multiple purposes.

2.2 Applications of sentiment analysis for real-world situations

A growing body of research aims at identifying whether sentiment mining provides commercial, economic or any other kind of value. The most common applications of sentiment analyses concentrate on predicting real-world outcomes or monitoring a brands online performance. This paper focusses on using sentiment analyses for monitoring purposes, nevetheless some examples from both applications are mentioned. For example, Bollen, Mao and Zeng (2011) provided a demonstration of how public moods on Twitter can actually improve the predictions of the Dow Jones Industrial Average (DIJA). Mishne and Glance (2006) found a significant correlation between movie mentions in weblogs and the movie's financial success. By conducting a sentiment analysis of the weblogs, the correlation between movie mentions and financial success was improved. Similarly to the study initiated by Mishne and Glance (2006), Asur and Huberman (2010) analysed both the number of mentions and the sentiment found on Twitter in order to predict movie box office revenues. Their study demonstrated that the forecasts from 3 million tweets outperformed the prediction accuracy of the Hollywood Stock Exchange (HSX). The authors also demonstrated the efficacy of sentiments present in tweets for improving predictions after a movie had been released. Research conducted by Tumasjan et al. (2010) investigated what the tweets on Twitter reveal about political sentiment and whether the messages reflect the election results. Their analysis indicated close correspondence of tweets mirroring the offline political sentiment. Tumasjan et al. (2010) also confirmed that the number of tweets mentioning a party reflects the election results. Pitt et al. (2011) collected sentiments and other social media data using Social Mention. With the data, the authors analysed the social media visibility of six wine brands using Chernoff faces. They conclude that using Chernoff faces to is a simple but powerful tool for understanding the complexity of a brand's social media presence. Likewise, Gamon et al. (2005) and Jansen et al. (2009) extracted sentiments about products or brands. They conclude that micro-blogging provides an environment in which customers speak their thoughts and opinions about a brand. Therefore, companies should explore the micro-blogging domain as part of their branding strategy. Furthermore, in a similar paper the same authors Jansen et al. (2009) investigated the real-time Twitter sentiment of numerous brands. Their conclusion is that, given the availability of social media monitoring tools, micro-blogging should be seen as competitive information source.

3. CONCEPTUAL MODEL

Figure 1 shows a model of comparison between the sentiment mining tools SocialMention and Semantria. Social Mention gathers input from 80+ different social media platforms, however, for my research I concentrated on the most popular microblogs which are Twitter, Facebook and Friendfeed. The input for Semantria can be freely chosen which means Semantria can be applied directly to any document or database of choice. In this study, a self-collected Twitter database, which I obtained from Twitter's application programming interface (API) served as the input for Semantria.

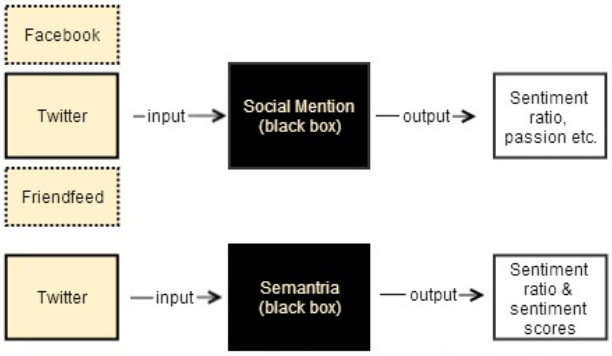

Figure 1: Model on sentiment mining tools

Next, the inputs are transferred to SocialMention and Semantria's black box. A black box is a system that is viewed in terms of the input it gathers and the output it generates. The internal workings of the transformation characteristics from input to output, the so-called 'black box', are not known. For sentiment mining tools, the black box consist of algorithms and uses sentiment lexicons in order to identify sentiment-bearing phrases. The output from the black box provides sentiment

[3] http://www.nltk.org/
[4] http://www.r-project.org/
[5] http://rapid-i.com/content/view/184/196/
[6] http://www.alias-i.com/lingpipe
[7] http://www.liwc.net/

indicators for a specific keyword. This means that different sentiment mining tools use a distinct algorithm (black box), which may influence the results of a sentiment analysis. In addition, Social Mention's black box also calculates a number of different metrics which are explained in a later section.

4. DATA COLLECTION

4.1 Data collection procedure

The data for this research were collected using ScraperWiki[8] and Social Mention. For both tools, the data collection was performed from May 1st – May 31st , 2014. The first step involved defining a keyword to which the tools gathered the available data. Following this step, both tools required different procedures. On ScraperWiki, the requested Twitter data was downloaded as a csv. or xls. file on a daily basis. This data was then cleansed and transferred to Semantria. In this context, cleansing means filtering the content to only show tweets in the English language. This was an important step, as the sentiment mining tool Semantria is only capable of conducting a sentiment analysis on content in the English language. Unlike ScraperWiki, Social Mention allows users to filter search queries according to language and time period with the click of a button. This allowed me to retrieve Social Mention data on a weekly basis, which was then written down. In addition, the data on Social Mention is presented as readily available data, therefore not requiring any additional steps of cleansing or filtering. During the data collection procedure I discovered inconsistencies in the data pertaining to the Lexus IS car model. On both data gathering tools, the 'IS' extension was often interpreted as an 'is', e.g. *"Lexus is great"*. Consequently, the Lexus IS car model was removed and does not feature in any of the results.

4.2 Twitter data

The dataset that I used was obtained from parsing regular feeds on Twitter[9]. This micro-blog is one of the most popular social media platforms, represents all ages and the majority of Twitter users are situated in the US (Leetaru et al., 2013). Furthermore, it allows users to easily mention anything from a brand to a person via the hash tag function. However, access to all of Twitter's data since its beginnings in 2006, its so called 'firehose', is only accessible with a significant research budget. Therefore, I am only able to use Twitter's API function, which grants access to a 1% random sample of the 'firehose' regarding a specific keyword. Nevertheless, my research involved gathering thousands of tweets for every keyword, which makes me believe that the data gathered from Twitter can be used as a representative sample. With the use of ScraperWiki, a web-based platform that extracts tweets from Twitter's API version 1.1, I obtained around 1% of all tweets related to a specified keyword from a single day. The web-based platform enabled me to receive the timestamp, language, ID, username and tweet text of thousands of relevant tweets for my analysis. However, inferences of age, gender and geographical location could not be made when analysing Twitter.

4.3 Social Mention data

Social Mention[10] is a social media search and analysis tool that aggregates user generated content from a number of social media platforms into a single stream of information. As it is one of the most popular tools and offers an overview of numerous metrics, this platform seems most appropriate to use. The Social Mention tool allows users to follow what and how much people are saying about a brand, product, service or indeed anything happening across the social media landscape. Platforms such as Twitter, Facebook, FriendFeed, Reddit and YouTube are only a few of the numerous online platforms from which Social Mention gathers user-generated content. Social Mention reports on a number of metrics with the most important defined in Table 1 below. However, for this research I have chosen to only concentrate on the sentiment and passion metrics. The reasoning behind this is that these two metrics are the only two that I believe provide a possible insight into people's opinions.

Table 1: Social Mention metrics: Descriptions

Metric	Definition	Algorithm/ How calculated
Sentiment	The ratio of generally positive mentions to the number of generally negative mentions.	Number of Generally Positive Mentions/ Number of Generally Negative Mentions This measure can also be gauged in absolute terms by counting the number of positive mentions, the number of neutral mentions and the number of negative mentions.
Passion	A measure of the likelihood of individuals talking about your brand in social media and will do so repeatedly.	A small number of individuals talking about a brand repeatedly will give a high passion score. A large number of individuals talking about your brand, but only infrequently per individual, will give a low passion score. Most frequently used keywords and number of times mentioned. Number of mentions by sentiment.
Strength	The likelihood that your brand is being discussed in social media.	Phrase mention within the last 24 hours divided by the number of total possible mentions.
Reach	A measure of the range of influence.	Ratio of the number of unique individuals talking about your brand as a % of the number of total possible mentions.
Unique authors	An indicator or the number of authors messaging about a brand.	The number of unique authors messaging about a brand within a particular time period.
Frequency	The frequency with which mentions of a brand appear.	Measured in minutes or seconds

With the use of Social Mention I collected various data on

[8] www.scraperwiki.com/
[9] https://twitter.com/

passion and sentiment scores for each of the 12 car models. The tool presents results, for a specific keyword e.g. *Honda Pilot*, from all of the platforms it gathers data from. However, the advanced filter options enable one to acquire any of the above mentioned metrics for a specific social media platform. Therefore, I gathered sentiment and passion scores of the three most popular micro-blogs on Social Mention, namely, Twitter, Facebook and FriendFeed. This allowed me to compare the scores and see whether there were any differences between the different social media platforms. The approach I used involved gathering sentiment and passion scores of weeks 1,2,3 and 4 in May.

5. METHOD

5.1 Sampling

The sample I chose consists of 12 car models (see Appendix A). Here I distinguished between standard and premium car models, and the three largest car-manufacturing areas in the world, namely North America, Europe and Asia. This means that four car models of each area represent two standard and two premium car models. To the best of my knowledge, the literature does not offer a methodology for optimal sampling size selection in the social media context. Consequently, it does not offer any methodology for which car models to choose. As a result, I chose 12 car models that are in the Top 250 list of car sales. As I had to individually collect and analyse vast amounts of social media data, a sample size of 12 car models seems plentiful.

5.2 Keywords

In order to collect Twitter and Social Mention data, it was necessary to specify keywords that identify mentions associated with each model. The easiest and best option was to choose the overall car model name, e.g. the Audi A6, rather than the Audi A6 3.0 TDI or Audi A6 2.0 T. This procedure was used for all of the car models except for the Volkswagen Passat and Volkswagen Tiguan. As tweets on Twitter are limited to 140 characters, the keywords Volkswagen Passat (17 characters) and Volkswagen Tiguan (17) are less popular as they exhaust large chunks of this limitation. More popular are the abbreviations VW Passat (9) and VW Tiguan (9) which allow users to post more about their opinions. As a result, these terms were used in this study.

5.3 Sentiment analysis

As mentioned in Section 2.2, there are a number of different methods and tools with which to conduct a sentiment analysis. This has resulted in no clear agreement as to which method and tool is best. For this paper I have chosen to use Semantria.

Semantria is a sentiment analysis solution created by Lexalytics Inc. [11], a well-known text analysis software provider. The sentiment software application Semantria offers a fee-based Excel plug-in that enables the analysis of Excel spreadsheets according to positive, neutral and negative sentiments. For my research, this kind of sentiment analysis was conducted on a dataset of 899 tweets per car model. The reason being that the fee-based tool Semantria offers a free trial which allows the user to analyse up to 10,000 document transactions. By multiplying the eleven car models by 899 tweets (on average 29 per day), I was able to make use of the free trial version. Nevertheless, 899 tweets per car model seems a large enough sample.

The Semantria Excel plug-in conducts an automated sentiment analysis of the dataset based on algorithms developed to extract sentiment in a similar manner as human beings According to Semantria[12], the extraction of sentiments in a document adheres to the following steps; (1) document broken into parts of speech (POS) tags, (2) algorithm identifies sentiment-bearing phrases, (3) logarithmic scale from -10 to 10 scores each sentiment-bearing phrase, (4) scores combined to determine overall sentiment. Through these statistical inferences, each tweet is tagged with a numerical sentiment value ranging from -2.0 to +2.0 and a polarity of (i) positive; (ii) neutral; or (iii) negative. Since its launch in 2011, a number of businesses and researchers have used Semantria to conduct sentiment analysis (Aston, Liddle & Hu, 2014; Abeywardena, 2014).

5.5 Limitations

Before continuing with the *Results* section, I highlight some limitations that should be considered for the rest of this paper. Firstly, Social Mention is a widely talked about social media analysis tool in the World Wide Web, however, its FAQ section[13] does not mention how it works and where it gathers its data from. One can only assume that it gathers data from a social media platform's API. This leads to a second consideration, which is that, if Social Mention accesses Twitter's API version 1.1, then the data should amount to the same as mine. However, whereas I was able to obtain thousands of weekly tweets for every car model, Social Mention presented me with far less data. In most cases the social media analysis tool provided between 20 to 250 per week. Lastly, I would like to highlight a further limitation in regards to Social Mention. As mentioned in Section 4.3, Social Mention offers a number of different metrics with short explanations how these are calculated. Therefore, I tested the proposed calculations to see whether I could generate the same results as Social Mention. My calculations were not able to replicate the same results for Social Mention's *passion*, *strength* and *reach* metrics. Consequently, I advise to treat the output from these metrics with a degree of caution. In addition, as Social Mention's *passion* metric could not be replicated, the initial idea of calculating the *passion* metric for Semantria's output and comparing the results with Social Mention's *passion* output was not possible. Because of this, I chose to focus my attention on the percentage of returning authors for both Semantria and Social Mention. Percentage of returning authors can be defined as the same author mentioning the same car model more than one time. I believe this measure is similar to Social Mention's passion indicator, which is defined as "the likelihood of individuals talking about your brand in social media and will do so repeatedly", albeit a more simplistic approach.

[11] http://www.lexalytics.com/

[12] https://semantria.com/features/sentiment-analysis

[13] http://socialmention.com/faq

6. RESULTS

6.1 Comparison of outputs for two sentiment mining tools

6.1.1 Comparison of sentiment ratios

Figures 2 and 3 show the car model sentiment ratios obtained from Semantria and Social Mention, respectively. The data for both sentiment mining tools were extracted from Twitter. This allows a comparison of the output for the two different sentiment mining tools. The x-axes represent the time, whereas the y-axes represent the sentiment ratios in whole numbers. These whole numbers denote positive-to-negative sentiment ratio, e.g. a value of 3 means a positive-to-negative ratio of 3:1. On two occasions a negative value can be observed, which means the negative sentiments outnumber the positive sentiment, e.g. 1:3. The coloured lines represent the sentiment ratios for the 11 car models that were used as the control variables for the two different sentiment mining tools.

In Figure 2, the Honda Accord and VW Passat show the lowest and most consistent sentiment ratios, whereas the Ford F-Series is the most consistent at the higher end of the sentiment ratios. The main abnormalities are found in the Acura TL's decline from 11:1 to 2:1 and the Audi A6's rise from 5:1 to 12:1, both happening from week 3 to week 4. Overall, the sentiment ratios for the eleven car models do not show a clear pattern and an exact range in which most of the sentiment ratios are located cannot be determined.

From Figure 3 we can observe that all but one car model, the Lexus ES, show positive sentiment ratios throughout. The Ford F-Series has the highest positive to negative sentiment ratio, however, from week 3 to 4 there is a large dip. At the bottom end of the graph, the Audi A6 shows the lowest but most consistent sentiment ratio. The overall majority of positive-to-negative sentiment ratios are situated between 4:1 and 7:1.

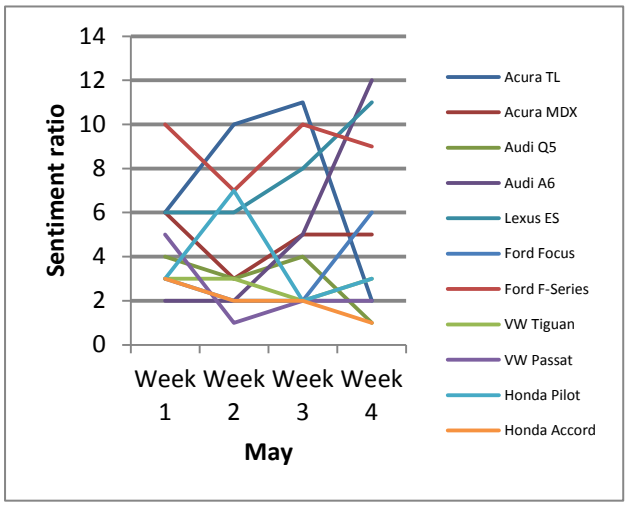

Figure 2: Semantria – Sentiment ratios for Twitter

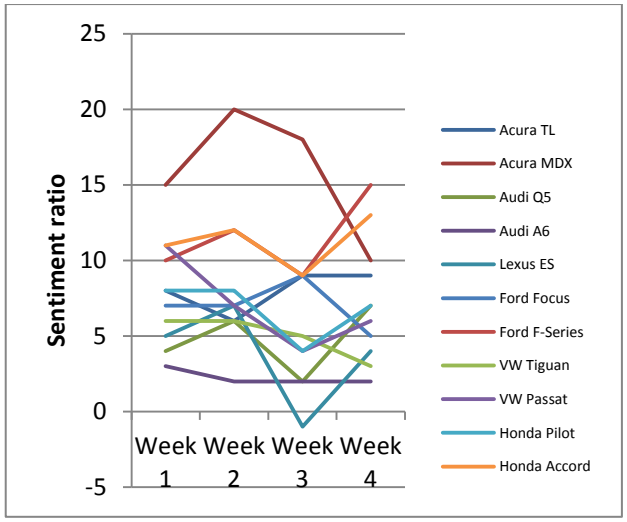

Figure 3: Social Mention - Sentiment ratios for Twitter

Tables 2 and 3 show the statistics and results of the paired samples t-test conducted for Semantria and SocialMention's sentiment ratios for Twitter. The sample size N=44 consists of the 11 car models for which the weekly (4 week period) data was gathered. The very small Sig. (2-tailed) value of 0.001 in Table 3 indicates that there is a statistically significant difference between the mean sentiment ratios for Semantria and Social Mention. Since Table 2 shows that Social Mention's mean sentiment ratio is 7.55 and Semantria's ratio is 4.66, I can conclude that the sentiment ratios obtained from Social Mention are significantly more positive than those obtained from Semantria.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Semantria	4.66	44	3.102	.468
	Social Mention	7.55	44	4.348	.655

Table 2: Semantria and Social Mention paired samples statistics box

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Semantria - Social Mention	-2.886	5.362	.808	-4.517	-1.256	-3.570	43	.001

Table 3: Semantria and Social Mention paired samples test

6.1.2 Comparison of returning authors

Figures 4 and 5 show the percentage of returning authors mentioning a specific car model over a four week period. Returning authors are the opposite of unique authors and are defined as the same author mentioning the same car model more than one time. The graphs show that, for both Semantria and Social Mention, the percentage of returning authors predominantly ranges from 20-45%, however, the distribution is very different. The output from Semantria shows that the percentage of returning authors is evenly spread for all the car models. The VW Passat has the overall highest percentage of returning authors, meaning that large percentage of same authors repeatedly mention the VW Passat in their tweets. At the lower end, both the Audi A6 and Audi Q5 show the lowest percentage of returning authors. In comparison to Fig. 4, the output from Social Mention in Fig. 5 shows a different distribution. For eight of the eleven car models, the majority

percentages of returning authors are located at around 40%. This means that according to Social Mention, the percentage of returning authors for most of the car models is higher and more consistent when compared to Semantria's output. Upon closer inspection, the comparison shows some large differences for some of the car models. For example, the VW Passat is ranked the highest according to the Semantria data, however, strangely enough for Social Mention the same car shows the lowest returning author percentage. Furthermore, the Audi A6 is among the lowest ranked in Fig. 4 yet ranks as one of the highest in Fig. 5. With the exception of the Audi Q5 and Acura TL, the output from both sentiment mining tools show no similarities between returning author percentages and the remaining car models. In summary therefore, it can be said that the distribution, rankings and fluctuations for the nine car models are very different for the two sentiment mining tools.

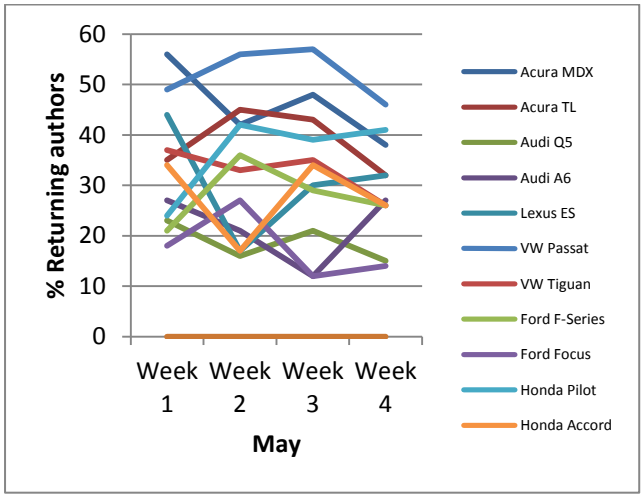

Figure 4: Semantria – Returning author percentages

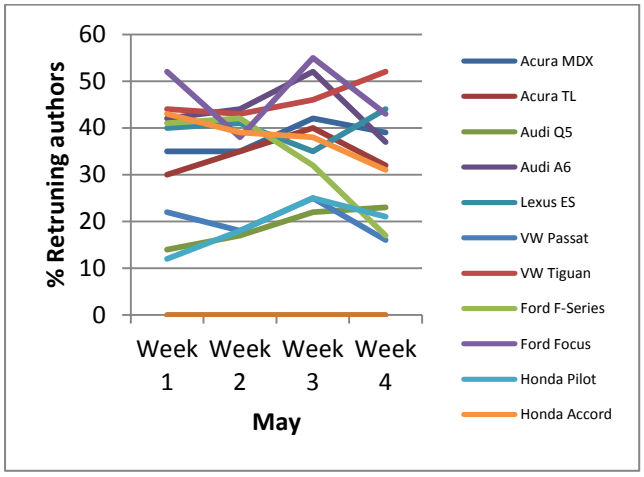

Figure 5: Social Mention - Returning author percentages

6.2 Comparison of sentiment and passion metrics for three different social media platforms

6.2.1 Comparison of sentiment ratios

Figures 3, 6 and 7 show the sentiment ratios obtained from Social Mention for the social media platforms Twitter, Facebook and Friendfeed..

The sentiment ratios for Twitter were explained in the previous section and I now move on to Figure 6 which represents data collected from Facebook. The graph shows that in Week 1 the sentiment ratios for more than half of the car models are at their highest, before they rapidly decline in Week 2. In general, the majority of models show an up, down, up, down pattern with large fluctuations in sentiment ratios. For example, the ratios for the Acura MDX are 25:1 in Week 1, 9:1 in Week 2, 17:1 in Week 3 and 12:1 in Week 4. Overall, the prevailing number of positive-to-negative sentiment ratios range between values of 5:1 and 11:1.

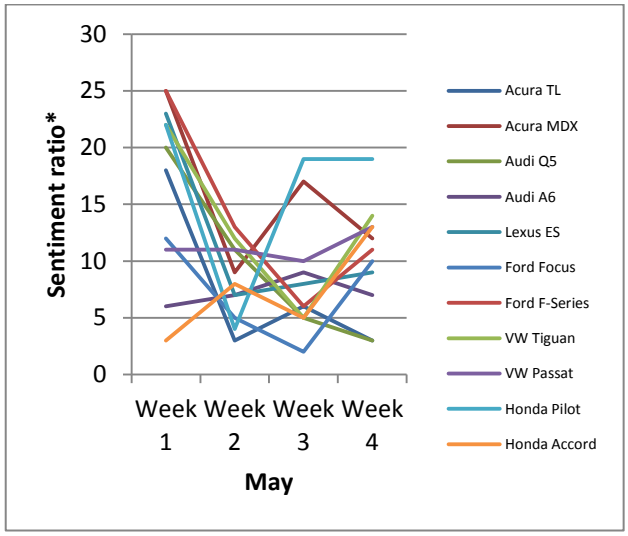

Figure 6: Social Mention - Sentiment ratios for Facebook

Lastly, Figure 6 shows the output of data collected from the social media platform Friendfeed. The graph depicts that the sentiment ratios for the majority of car models remain relatively low but constant. Notable outliers are the Week 1 and Week 4 values for the Honda Accord, and the negative ratio in Week 4 for the Acura TL. Furthermore, some car models show sentiment ratios of 0, meaning there were no positive or negative mentions. In these cases all of the mentions were neutral. The prevailing number sentiment ratios range from values of 0:0 to 4:1.

The results of two paired t-test and one Wilcoxon signed rank test show that there are significant differences in sentiment ratios for all three platforms (see Appendix B). The low Sig. (2-tailed) value of 0.003 (see Appendix B1) signals that there is a significant difference between sentiment ratios on Twitter and Facebook. Twitter has a mean sentiment ratio of 7.55, whereas Facebook has a mean ratio of 10.98. Therefore it can be said that the sentiment ratio on Facebook is significantly more positive than on Twitter. To be exact the sentiment ratio on Facebook is an average of 3.42 higher.

The comparison of sentiment ratios for Facebook and Friendfeed shows a very low Sig. (2-tailed) value of 0.000 (Appendix B2). This signals a very significant difference, which can also be observed from the large difference in mean sentiment ratios for Facebook (10.98) and Friendfeed (2.89).

Due to reasons pertaining to not all assumptions and conditions being fulfilled, the significance of sentiment ratio differences for Twitter and Friendfeed was carried out using Wilcoxon's signed rank test. The test shows a further very low Sig. (2-tailed) value of 0.000 (Appendix B3). This means that a significant difference in mean sentiment ratios between Twitter and Friendfeed exists.

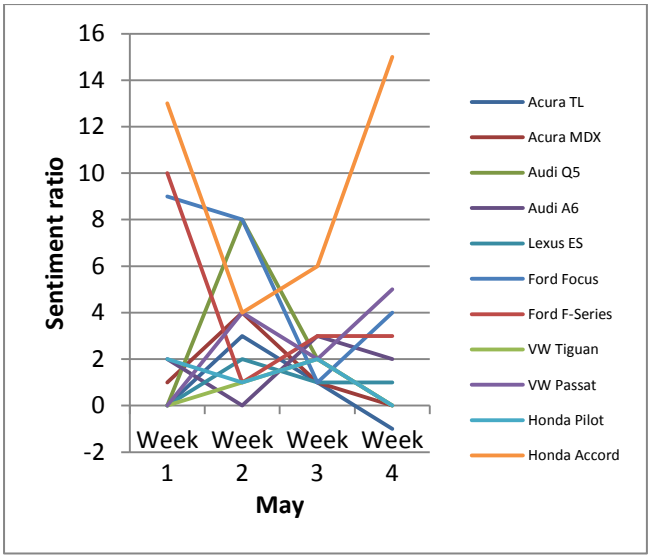

Figure 7: Social Mention - Sentiment ratios for Friendfeed

6.2.2 Comparison of passion scores

Following a look at the differences in sentiment ratios, this section provides an overview in differences of passion scores for Twitter, Facebook and Friendfeed. The data which allows this comparison was extracted from Social Mention's *passion* metric and is presented below (see Appendix B4). From this table several observations can be made. Firstly, the passion for the same car model on different social media platforms shows no consistencies. Secondly, the majority of car models, with the exception of the Acura MDX, show their lowest passion scores on Facebook. In some cases, Facebook also shows a passion score of 0, indicating that in some weeks there is no likelihood of users repeatedly mentioning the same car model. Interestingly, the passion for some car models is very high on Twitter and Friendfeed. This means that users are more likely to repeatedly mention the same car model again on these social media platforms. The results show that the car models users are most passionate about are the Lexus ES, VW Passat and Honda Pilot. The Audi Q5 and VW Tiguan show the least passionate users.

7. DISCUSSION

7.1 Interpretation of results

The results of this study show that there are statistically significant as well as observational differences in the outputs for different sentiment mining tools and social media platforms. This is mainly attributable to the algorithms, so-called black boxes, being used by the different tools. Nevertheless, some inferences can be made as to why this might be the case, by analysing a car model's social media presence. An overview of each car model and car brands social media presence according to their followers and likes on Twitter and Facebook is provided in Appendix C. The bold numbers refer to unofficial fan-pages, whereas the remaining numbers refer to official fan-pages that are being managed by the respective car brands. From the table we can see that all car brands have official fan-pages on Twitter and Facebook, the leaders of both belonging to Audi USA. However, only Ford and Honda manage official fan-pages for their range of car models. The Ford Focus, Ford F-Series (Trucks), Honda Pilot and Honda Accord all show high numbers of followers and likes. Because of this, one might expect to see higher sentiment ratios for these car models. However, in Figures 3 and 6 the sentiment ratios for these car models do not stand out, they are merely on par with the other car models that do not boast any official fan-page. With this interpretation it is fair to assume that brands engaged in a full social media campaign are able to generate a great number of followers, but are also susceptible to more negative attention. A further assumption could be that social media campaigns do not have any bearing on the opinions users post online.

Moving on to the interpretation of the data for different social media platforms, I can only assume the reasons behind some of the results are attributed to the different demographics of these platforms. One of those assumptions concerns the sentiment ratios obtained from Friendfeed. In some cases the sentiment ratio is 0:0, meaning there were no positive or negative comments being made. The reason behind this could be attributed to the demographics of Friendfeed, as this platform is not as popular as Twitter and Facebook. Consequently, this led to only a small number of mentions. As most online opinions are found to be neutral, the small number of mentions limited the chances of positive or negative comments. A further assumption relates to the low passion scores obtained from Facebook and can be explained by Social Mention's definition of passion, *"A small number of individuals talking about a brand repeatedly will give a high passion score. A large number of individuals talking about your brand, but only infrequently perindividual, will give a low passion score"*. It is widely documented that Facebook boasts the most number of users, therefore the likelihood for this large number of users to talk infrequently about a brand is higher and could be the reason behind the low passion score. Conversely, the smaller number of users on Friendfeed could be the reason for the high passion scores found on this platform.

7.2 Limitations

There are several limitations that provided some obstacles during the research I conducted. In Section 5.5, I highlighted the issues concerning Social Mention. In this Section I would like to mention some limitations regarding the interpretation of the results, i.e. explaining links between the data. This is a task I found difficult, as the data does not provide clear patterns. I was expecting the results to be more consistent across the different tools and platforms, however, there are some huge discrepancies. For example, according to the Semantria data, the Honda Accord has a very low sentiment ratio, conversely the data from Social Mention shows it has a relatively high sentiment ratio. This high sentiment ratio would lead me to believe that the passion for the Honda Accord is also very high; however, this is not the case. This problem I encountered with interpreting the relationships between the data, is something I believe would be of interest for future research. It would be interesting to investigate how the different metrics are associated with each other.

8. CONCLUSION

Because sentiment mining is growing in popularity, it is important to know how sentiment ratios for different sentiment mining tools vary for the same sample. This paper has reviewed two sentiment mining tools and concludes that there is a significant difference in sentiment ratio outputs for Semantria and Social Mention. Overall, the sentiment ratios obtained from Social Mention were more positive than those from Semantria. In addition, an analysis of sentiment ratios and passion scores was also conducted. From the Social Mention analysis conducted on Twitter, Facebook and Friendfeed, I can conclude that the outputs shows statistical and observational differences. Overall, the sentiment ratios on Facebook are the highest for all three platforms. By far the lowest sentiment ratios can be found on Friendfeed and the sentiment ratios for Twitter are situated

inbetween. As for the passion scores, my observations show that Friendfeed and Twitter have relatively high and in some cases similar passion scores, whereas Facebook shows relatively low and even non-existent passion scores. The consensus that can be drawn from these results is that sentiment mining tools do indeed show reliability issues as they are influenced by a number of different variables.

I believe that the value of this study has both theoretical as well as practical implications. Firstly, there is currently no scientific literature available on the reliability of sentiment mining tools and this paper aims to fill this void by comparing Semantria and Social Mention. Secondly, this paper provides an overview of how different brands are presented on different social media platforms. Lastly, although sentiment mining delivers insights into what customers think about products, services and brands, the information of these insights should be used with caution, as sentiment- as well as social media-tools are prone to reliability issues.

ACKNOWLEDGEMENTS

I would like to thank Dr. A.B.J.M. (Fons) Wijnhoven and an anonymous reviewer for their constructive feedback throughout the whole process.

9. REFERENCES

Abeywardena, I. S. Public Opinion on OER and MOOC: A Sentiment Analysis of Twitter Data. *Proceedings of the International Conference on Open and Flexible Education (ICOFE 2014), Hong Kong SAR, China.*

Aston, N., Liddle, J., & Hu, W. (2014). Twitter Sentiment in Data Streams with Perceptron. *Journal of Computer and Communications, 2014*.

Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on* (Vol. 1, pp. 492-499). IEEE.

Bollen, J., Mao, H., Zeng, X., (2011). Twitter mood predicts the stock market, *Journal of Computational Science 2*, pp.1-8

Choi, H., and Varian, H. (2011). "Predicting the Present with Google Trends," working paper.

Davidov, D., Tsur, O., & Rappoport, A. (2010, August). Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 241-249). Association for Computational Linguistics.

Du, R. Y., and Kamakura, W. A. (2012). "Quantitative Trendspotting," *Journal of Marketing Research* (49:4), pp. 514-536

Esuli, A., and Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC.*

Fiaidhi, J., Mohammed, O., Mohammed, S., Fong, S., & hoon Kim, T. (2012, August). Opinion mining over twitterspace: Classifying tweets programmatically using the R approach. In *Digital Information Management (ICDIM), 2012 Seventh International Conference on* (pp. 313-319). IEEE.

Gamon, M., Aue, A., Corston-Oliver, S., & Ringger, E. (2005). Pulse: Mining customer opinions from free text. In *Advances in Intelligent Data Analysis VI*(pp. 121-132). Springer Berlin Heidelberg.

Go, A., Huang, L., & Bhayani, R. (2009). Twitter sentiment analysis. *Entropy,17.*

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, *60*(11), 2169-2188.

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009, April). Micro-blogging as online word of mouth branding. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems* (pp. 3859-3864). ACM.

Kim, S. M., & Hovy, E. (2004, August). Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics* (p. 1367). Association for Computational Linguistics

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg!. *ICWSM*, *11*, 538-541.

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., & Shook, E. (2013). Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, *18*(5).

Nasukawa, T., Yi, J. (2003). Sentiment Analysis: Capturing Favorability Using Natural Language Processing, *K-CAP'03, October 23–25, 2003, Sanibel Island, Florida, USA.*

Copyright 2003 ACM 1-58113-583-1/03/0010

Mishne, G., & Glance, N. S. (2006, March). Predicting Movie Sales from Blogger Sentiment. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (pp. 155-158).

Ostrowski, D. A. (2010, September). Sentiment mining within social media for topic identification. In *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on* (pp. 394-401). IEEE.

Pak, A., and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proc. of LREC*.

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

Pitt, L., Mills, A. J., Chan, A., Menguc, B., & Plangger, K. (2011, June). Using Chernoff faces to portray social media wine brand images. In *6th AWBR International Conference, Bordeaux, France*.

Read, J. (2005, June). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop* (pp. 43-48). Association for Computational Linguistics.

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.

Thelwall, M., Wouters, P., & Fry, J. (2008). Information-centered research for large-scale analyses of new information sources. *Journal of the American Society for Information Science and Technology*, *59*(9), 1523-1527.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *ICWSM*, *10*, 178-185.

Wilson, T., Wiebe, J., Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pp. 347-354, Vancouver, October 2005

Yu, H., & Hatzivassiloglou, V. (2003, July). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* (pp. 129-136). Association for Computational Linguistics.

Zachariadis, M., Scott, S., Barrett, M. (2013). Methodological Implications of Critical Realism for Mixed-Methods Research. *MIS Quarterly Vol. 37 No.3, pp. 855-879/September 2013*

APPENDIX A - SAMPLE

Car Model	Keyword	Grouping by Standard/Premium	Country
Volkswagen Passat	VW Passat	Standard	Europe
Volkswagen Tiguan	VW Tiguan	Standard	Europe
Honda Accord	Honda Accord	Standard	Asia
Honda Pilot	Honda Pilot	Standard	Asia
Ford Focus	Ford Focus	Standard	North America
Ford F-Series	Ford F-Series	Standard	North America
Audi A6	Audi A6	Premium	Europe
Audi Q5	Audi Q5	Premium	Europe
Lexus ES	Lexus ES	Premium	Asia
~~Lexus IS~~	~~Lexus IS~~	~~Premium~~	~~Asia~~
Acura MDX	Acura MDX	Premium	North America
Acura TL	Acura TL	Premium	North America

APPENDIX B – OUTPUT STATISTICAL TESTS AND PASSION TABLE

B1 – Twitter and Facebook: Paired samples t-test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Twitter	7.55	44	4.348	.655
	Facebook	10.98	44	6.436	.970

Paired Samples Test

	Paired Differences							
				95% Confidence Interval of the Difference				
	Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1 Twitter - Facebook	-3.432	7.235	1.091	-5.631	-1.232	-3.146	43	.003

B2 – Facebook and Friendfeed: Paired samples t-test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Facebook	10.98	44	6.436	.970
	Friendfeed	2.89	44	3.539	.533

Paired Samples Test

	Paired Differences							
				95% Confidence Interval of the Difference				
	Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1 Facebook - Friendfeed	8.091	7.600	1.146	5.780	10.401	7.062	43	.000

B3 – Twitter and Friendfeed: Wilcoxon signed rank test

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Twitter	44	7.55	4.348	-1	20
Friendfeed	44	2.89	3.539	0	15

Test Statistics[a]

	Friendfeed - Twitter
Z	-4.894[b]
Asymp. Sig. (2-tailed)	.000

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

B4 – Passion % for Twitter, Facebook and Friendfeed

Car Models	Twitter				Facebook				Friendfeed			
	Week 1	Week 2	Week 3	Week 4	Week 1	Week 2	Week 3	Week 4	Week 1	Week 2	Week 3	Week 4
Acura MDX	21	51	32	21	24	12	37	36	20	65	11	14
Acura TL	40	22	11	25	16	4	0	4	14	57	28	25
Audi Q5	8	8	7	26	12	8	8	21	40	43	14	28
Audi A6	21	24	22	29	26	13	0	4	22	34	30	23
Lexus ES	44	38	33	53	25	0	0	8	25	20	40	37
VW Passat	43	38	20	25	20	16	8	16	57	50	58	41
VW Tiguan	17	21	28	35	9	26	14	40	14	25	16	20
Ford F-Series	27	31	29	16	28	0	10	5	43	33	16	25
Ford Focus	31	44	52	38	4	4	0	4	53	21	37	40
Honda Pilot	29	25	30	15	16	0	16	4	28	63	25	40
Honda Accord	15	16	13	10	4	0	8	4	55	46	61	35

APPENDIX C – CAR MODEL AND CAR BRAND SOCIAL MEDIA PRESENCE

Car models	Twitter	Facebook
Acura	58,900	601,477
Acura TL	0	**10,311***
Acura MDX	0	**3,689**
Audi USA	743,000	8,661,554
Audi Q5	0	**218,145**
Audi A6	0	**101,287**
Lexus	468,000	3,056,504
Lexus ES	0	**1,873**
Ford	437,000	2,535,230
Ford Focus	18,300	936,373

Ford F-Series (Trucks)	100,000	1,316,248
VW USA	265,000	2,800,334
VW Tiguan	0	**11,889**
VW Passat	**1,236**	**2,887**
Honda	308,000	3,312,781
Honda Pilot	0	51,636
Honda Accord	0	727,318

*****The bold numbers refer to either unofficial groups or Wikipedia entries that are followed and liked by users. These fan-pages are independent of each other and are not managed or endorsed by any of the abovementioned brands.
The data were gathered on the 31st of May.