ABSTRACT
Users of social media increasingly express their opinions and as a result, a large number of commercially and freely available sentiment analysis tools have been developed. These tools claim that they can measure the sentiment expressions on anything, but their validity and reliability has not yet been determined. This research focuses on the concurrent validity of various sentiment analysis tools to determine if these claims can be corroborated. In addition, the consistency of their algorithms is also researched. Four of the five tools that were tested appear to have more validity as well as reliability than the fifth tool. This observation can be explained by looking at the analytic procedure of the tools. The results have theoretical as well as practical consequences for academics researching sentiment analysis as this research shows the effect of various analytics procedures. Another practical contribution is that users of the tools, researched in this study, have (more) proof of the accuracy of the tool they use. These users often take the sentiment reports into account in their decision making, thus making this proof appreciated.

Keywords
Sentiment analysis, tool, concurrent validity, social media, consistency, algorithm, Text2data, Semantria, Meaningcloud, Sentirate, Umigon

1. INTRODUCTION
With the rise of Web 2.0 and the following growth of microblogging platforms, the number of people that express their opinion about certain products or services is increasing [4]. This customer feedback, if yielded correctly, is interesting for the decision-making of manufacturing companies of these products [10]. Not only is this trend visible on review sites designed for this purpose, but also at social media sites such as Facebook and Twitter. The content of this publicly available content may contain a lot of information regarding user experience. In addition, the marketing department can benefit from this insight as 81% of (American) internet-users state they have done online research on a product at least once [17]. Among these readers of online reviews, between 73% and 87% claim their purchase was significantly influenced by these reviews [17].

Since there are likely to be hundreds or even thousands of these reviews per product, the analysis cannot be performed manually, but must be done automatically. The field of sentiment analysis covers this topic. Plenty of research about the classification of sentiment expressions has already been conducted [1, 16, 21]. This classification determines that words such as “alright” and “sufficient” are weak positive sentiment indicators in a text whereas words as “sucks” and “hate” are strong negative sentiment indicators. In this fashion, a classification algorithm gives a polarity, often accompanied with a score to certain words to determine the overall sentiment of a text. The main focus of these researches is to improve this classification algorithm, where efforts are made to identify issues such as sarcasm in a text.

1.1 Problem Statement
The existing research on sentiment analysis is interesting, but very theoretical. Together with the growth of awareness of this subject and the value of information that can be obtained with it, many tools have been developed that are both commercially and freely available to perform sentiment analysis. Since these tools are increasingly used by companies in decision-making [6] it is worth asking if they are valid and reliable. Validity in empirical research broadly concerns that the method measures what it intends to measure, whereas reliability concerns consistency found in repeated measurements of the same phenomenon [8].

One of the many proofs that a measurement A is valid is if its results have a high correlation to the results of another measurement B that is intended to measure the same variable. Achieving this kind of validity makes results more credible. This validity is called concurrent validity, which is an aspect of criterion validity with the other measurement being the criterion [8]. So far there has been no research to my knowledge about this correlation between the results of different social media sentiment analysis tools running on the same query. Some tools offer the opportunity to sample social media data themselves, which can be analyzed for sentiment by the tool consequently. It is unlikely that different tools get the exact same dataset from running on the same query. To solve this problem it is necessary to gather a social media dataset myself and load these into the sentiment analysis tools. The sentiment results of the various tools on this dataset can be compared to check for concurrent validity of the algorithm of these tools.

Furthermore, it would be interesting to see how reliable the social media sentiment analysis tools perform. One approach to estimate reliability is to measure the “concurrent” consistency. This can be researched by letting the various tools analyze multiple datasets and looking at how the results from dataset pairs differ for each tool. The average differences for the
dataset pairs can consequently be calculated and the consistency can be determined by comparing the tool’s difference to the average difference for the dataset pairs.

1.2 Research Questions

Proceeding from the problem statement, this research addresses the following research questions:

- How do sentiment ratios vary from various sentiment analysis tools running on the same dataset containing social media data?
- For these sentiment analysis tools, how consistent are their performances?

The first question concerns the concurrent validity of the algorithms used by the tools, where the second question concerns their consistency.

These questions intentionally miss some specification (what sentiment analysis tools, social media, dataset collection methods and search queries will be used). These choices and their argumentation are discussed in Section 3. Prior to this section background information regarding sentiment analysis and validity is given in Section 2. The results of the comparisons between the various sentiment analysis tools are presented in Section 4. Ultimately I discuss the interpretation of my research as well as its limitations in Section 5 and conclude the paper in Section 6.

2. BACKGROUND

2.1 Related work

The contributions of data-driven research are that it provides greater sample sizes and unremitting evaluation of scientific theories, but on the other hand encourages passive data collection without experimenting and testing [11]. This paper is a data-driven research on sentiment analysis, but focuses on doing exactly what is lacking, which is testing.

Besides the theoretical research on sentiment analysis mentioned in the introduction, there has been research on the use of it for practical situations. The existing practical research differences from this research as it focuses mainly on predicting outcomes. An interesting article has been written about predicting election results using sentiments on Twitter [20]. It concluded that the results of analyzing Twitter data came close to traditional election polls. However, another research using the same technique in a different election concluded that those results were not repeatable [12]. The authors argued that political conversation is to be analyzed differently and due to indecisive voters plus spam posts on social media the predictions were not accurate. Another predicting research has been conducted regarding movies [5]. The aim was to predict the revenue of movies in the opening weekend according to the sentiments expressed on Twitter. The conclusion was that their predictions outperformed the Hollywood Stock Exchange which is considered as a golden standard in the industry. One article has compared 52 studies regarding predicting outcomes using social media and made two observations; (1) Sophisticated search term selection is of great importance in gathering microblog data, (2) using different sentiment analysis approaches results in controversial outcomes, which calls for a refined approach to analyze social media data [13].

What still lacks is research on actual sentiment analysis tools, either commercially or freely available and the validity and reliability of their results in particular. There has been research on the external validity of sentiment mining reports [22]. This research identified threats to external validity and conceptualized its dimensions, but does not concentrate on particular tools.

2.2 Validity and reliability

This research focuses on the concurrent validity (which is a component of criterion validity) and consistency of sentiment analysis tools. There are, however, more types of validity and reliability which are discussed in this section. Alongside with the theory of the various types of validity and reliability I discuss ways in how they can be used for researching sentiment analysis tools. This could be used for future research which is highly needed in this field.

2.2.1 Validity

As mentioned in Section 1.1, a method is test valid in general if it measures what it intends to measure. While this definition seems straightforward, the various types of validity take a different approach to determining the extent to what a method measures what it claims to. Test validity consists of content, construct and criterion validity.

Content validity involves ‘the systematic examination of the test content to determine whether it covers a representative sample of the behavior domain to be measured’ [3]. For this the algorithm used by the sentiment analysis tool could be researched to determine whether it does actually cover the sentiment expressions used in texts. Construct validity refers to ‘the extent to which the construct of theoretical interest are successfully operationalized in the research’ [2]. For sentiment analysis tools research these constructs could be the attributes of a product. Sentiments can for example be about the looks, the cost, user friendliness and efficiency of a product. An interesting research could be to see if analyzing these independent attributes combined yields the same sentiment result for a product as analyzing the same product in general.

Apart from test validity there is experimental validity; one section of this is external validity, which is concerned with generalization. Primarily to the extent that findings of a research can be generalized to other measures, populations, settings and times [18].

All the aspects of test validity of sentiment analysis tools are in need of research. The reason I chose concurrent validity is because I believe that it is the most feasible option for the available time frame of this research.

2.2.2 Reliability

As mentioned in Section 1.1, this research also focuses on ‘concurrent’ consistency. I put concurrent in quotation marks as this is not an official term. I did this to distinguish this form of consistency from internal consistency, which is another method to determine reliability. Internal consistency measures the correlations between various items in the same test [8]. This can be applied to sentiment analysis research by letting a tool analyze a dataset containing sentences with inverted meaning (e.g. “I love riding my bicycle in the rain” and “Cycling in the rain is the
worst”). The tool would have a high internal consistency rate if it gives a positive sentiment to the first sentence and a negative sentiment to the latter. Internal consistency differs from concurrent consistency whereas the first concentrates on only one test, the latter determines reliability over multiple tests. Also the concurrent consistency is measured as a comparison to the other tools, or rather the average of the other tools.

Another, and perhaps the easiest, method to measure reliability is test-retest. The same test is conducted after a period of time to see if the results are still the same [8]. This can be done for sentiment analysis by letting a tool analyze the same dataset after a period of time. Given that the tool has not been subject to any updates, the results should be exactly the same.

Concurrent consistency has been chosen as the method to estimate reliability of the tools. This is done because I believe it fits best with the approach of this research. As I collect social media data where I have no influence on the content, I most likely cannot check for internal consistency. Test-retest seems superfluous with this kind of software research, but can be checked in future research to be sure. Concurrent consistency can be determined, because I collect multiple datasets which will all be analyzed by various tools. The exact procedure for calculating the consistency follows in Section 4.3.

3. METHOD

3.1 Social media

With Facebook1 undoubtedly being the largest social media today (with 67% of the internet users having an account in 2012) [9], it is still important to ascertain that this is the best option for this research. Other popular social media sites include Twitter2, Pinterest3, Instagram4 and Tumblr5 [9]. Only the first on this list, however, is text based (the others are image based), which is crucial for this research. Extracting data from Facebook statuses could be problematic regarding privacy issues. There are sentiment analysis tools (such as SocialMention6 and Trackur7) that claim they can extract Facebook data, but the first impression is that this regards public pages and not personal statuses. On the other hand, Twitter data (also known as a collection of tweets) is gathered relatively easy. There is a 140 character limit that Twitter imposes on the messages of its users which often forces the users to abbreviate words or use slang in their sentiment that is difficult to analyze [14]. If all tools have to work with this disadvantage then this variable remains the same. Concluding, the source of the social media data that will be analyzed by a sentiment analysis tool in this research is Twitter.

3.2 Sentiment Analysis Tools

When looking for sentiment analysis tools online, one can conclude that there are tons of options available. These will probably all differ in their algorithm as well as their practical way of operating. Adequate sentiment analysis tools for this research must meet three requirements. The first requirement concerns the availability of the tool. Most sentiment analysis tools are commercially available and offer some form of trial version. This trial version should have the same functionality as the paid version and only differ in the quantity of query calls that can be made. In addition to this requirement, it is important that the tools are (relatively) popular in usage. This way, the research contributes to the practical aspect of sentiment analysis research. The last requirement for the chosen tools is that they have the availability to load a pre-made dataset. This is needed to research the concurrent validity and reliability of the sentiment analysis algorithm that they use.

This initial exploratory research resulted in testing dozens of tools as there is no scientific literature available that compares sentiment analysis tools. A categorization of the tools was observed, tools could either only analyze sentiment of their own sampled social media data or only analyze sentiment of external gathered data provided by the user. I have not found tools that could do both. This may be useful knowledge for future research regarding social media sentiment analysis tools.

The second observation regards the format of the data one wants analyzed by sentiment analysis tools. The majority of the tools either have a Microsoft Excel add-in or a demo version working on their website. This observation consequently means that it is most practical if my Twitter dataset is in text or .csv format. In the following subsections I briefly describe the five tools that met previously set requirements.

3.2.1 Text2data

Text2data is a start-up offering text analytics software as a service. Not only does it offer a Microsoft Excel add-in for sentiment analysis, but the tool has also specialized in Twitter-like content. They describe this function as a “more accurate analysis for short content (including irony) plus slang and abbreviation detection.” Their trial offers a credit of 300 transactions to be used in a month. Unfortunately, their Microsoft Excel add-in could not be installed due to a software error. However, with the use of the demo on their website8, it is possible to use these 300 document transactions manually. They operate by combining four approaches to sentiment analysis: (1) keyword spotting, (2) lexical affinity, (3) statistical methods, (4) concept-level techniques. Keyword spotting involves looking for unambiguous sentiment classifying words such as “happy” and “boring” [7]. This approach easily looks over negations and only looks at the most obvious words. Lexical affinity is slightly more sophisticated as it assigns a probability to unpredictable words [7]. For example, “accident” might be assigned a 75% chance of being a negative sentiment as it can not only be used negatively but also in the neutral expression “by accident”. Statistical methods are machine learning methods where the algorithm is being improved with a hand classified text [7]. Concept-level techniques involve looking semantic features where words are analyzed by their context [7]. “Long” can for example be positive when it’s about the longevity of a product but negative when it’s about a waiting queue. All these approaches combined result in a document score ranging between -1.0 and +1.0 as well as a default polarity tag that is negative, neutral or positive.

---

1 http://facebook.com
2 http://twitter.com
3 http://pinterest.com
4 http://instagram.com
5 http://tumblr.com
6 http://socialmention.com
7 http://trackur.com
8 http://text2data.org/twitter-sentiment-engine/
9 http://text2data.org/Demo/Demo
3.2.2 Semantria
Semantria is a text analytics tool developed by Lexalytics\textsuperscript{10}. Not only does it offer sentiment analysis, but also entity extraction, categorization and clustering. It claims to be "the world's largest cloud text mining service, analyzing over 8 billion documents per month"\textsuperscript{11}. This would explain why it is top-ranked when searching for sentiment analysis tools using a search engine. They offer a free demo of their Microsoft Excel add-in which allows the user to analyze up to 20,000 document transactions. Semantria’s sentiment analysis is based on Natural Language Processing where a sophisticated algorithm extracts the sentiment from a document. This procedure is carried out in four steps; (1) the document is broken into parts, (2) algorithms identify sentiment-bearing phrases, (3) these phrases earn scores, (4) scores are combined to an overall document score ranging between -2.0 and +2.0 and a corresponding default polarity tag\textsuperscript{12}.

3.2.3 Meaningcloud
Meaningcloud is a commercial tool originating from Spain that classifies and analyzes unstructured data. It offers a Microsoft Excel add-in with 40,000 requests in the free plan. A request is equal to an analysis of any text up to 500 words\textsuperscript{13}. The tool firstly discriminates opinions from facts and then combines the polarity of different sentences for a polarity of the whole text analyzed. The result is a polarity tag (P+, P, NEU, N or N-) combined with a polarity score (ranging between -1.0 and +1.0).

3.2.4 Sentirate
Sentirate is a service that offers sentiment analysis to any digital text form. It has a trial functionality integrated in their website that can be used before one potentially buys an API plan. This results in manual input for this research, just as with Text2Data. Sentirate does not state what algorithm they use nor how the analyzing procedure works. The result is a sentiment score ranging from -1.0 to 1.0 alongside a polarity tag (negative, fairly negative, a little negative, neutral, a little positive, fairly positive, or positive).

3.2.5 Umigon
Umigon is a freely available web application that detects sentiment in Twitter data\textsuperscript{14}. It is developed by an assistant professor and is completely open source. The operation consists of four steps; (1) detection of semantics in entire tweet, while also looking at smileys, (2) evaluation of hashtags, (3) decomposition of the tweet into a list of unigrams, bigrams, trigrams and quadrigrams to be compared with the terms in lexicons and sentiments are given for each match, (4) a last look at the entire tweet and giving the tweet a final polarity. Umigon does not give scores like the other tools, only a default polarity tag.

3.3 Products for search queries
In theory, every type of product could be chosen for this research. However, for this research there are a few requirements to the type of product that would make it more useful. The most important requirement is that these products are being mentioned often in social media, as this increases the amount of data that is available to analyze. This results in a higher credibility that the ratings truly reflect the population’s opinion about a product. As the quantity of the opinion rise, the chances decline that the average opinion of the product is not representative for the average sentiment. In simpler terms, a product that is popular in social media mentions is desirable. It is assumed that there is a correlation between amount of users and social media mentions. This consequently means that the second requirement for the product type is that it is used by the general population and not just a subgroup. Also contributing to this popularity is the social media user’s sense that his/her opinion matters and has some value. It is assumed that the higher the cost of a product, the more likely people search online for reviews before making a decision and hereby probably valuing these opinions on social media more. Thus, an assumption is made that more expensive type of products that are used by the general population are mentioned more in social media. The last requirement to be added is that the products have new versions coming out over the years. This has been recommended for sentiment analysis tools research as “this will allow for more comprehensive testing of the influence of events”\textsuperscript{15}.

With all these requirements in mind, I have chosen smartphone models as the product type to research. In June 2013, 56\% of American adults owned a smartphone and this percentage will keep growing, research indicated\textsuperscript{16}. It also states, that current owners of smartphones are likely to buy a new model when they feel that their phone is lacking behind, or when they are allowed by their service provider. For the queries of the research questions I choose various phones, to exclude the possibility that the results are affected by using only one smartphone model. This is the list of the smartphones that will be used in the steps mentioned in Section 3.4:

- iPhone 6 Plus
- Samsung Galaxy S5
- Nexus 6
- Sony Xperia Z3

Because of their popularity and recent release it most likely results in more actual data. The reason that I choose the Plus version instead of the default iPhone 6 is that when mining for the “iPhone 6”, the sentiments about the “iPhone 6 Plus” would also show up. A side note here should be added that there are multiple models of the phones, which have the same technical specifications but differ in cellular connectivity depending on the country and service provider. An assumption is made that these model numbers will not be mentioned by the majority of the Twitter population.

3.4 Data collection
Taking in mind that the trial version of Text2data is limited to 300 requests each dataset can maximally consist of 75 tweets (as there are four smartphone search queries). This led me to using Docteur Tweety\textsuperscript{17}, a service for exporting twitter data to a Microsoft Excel file. Their pricing depends on the quantity of tweets exported, but the last 50 elements per search can be

\textsuperscript{10} http://lexalytics.com
\textsuperscript{11} https://semantria.com/about
\textsuperscript{12} https://semantria.com/support/resources/technology
\textsuperscript{13} https://www.meaningcloud.com/developer/account/licenses/plan/Free
\textsuperscript{14} http://www.docteur-tweety.com/twexlist/?lang=en
exported for free. These Excel datasets can be acquired by filling in a twitter address link for the application to export. These twitter links were obtained in the following way; (1) Going to the Advanced Search of Twitter\(^\text{15}\), (2) Filling in a smartphone name in the box “this exact phrase”, (3) Setting the language to English. When searching for an exact phrase, only tweets that have “everything in between” will come up instead of sentences that have fragments such as “everything” and “between” separated. The last step is taken to prevent the sentiment analysis tools of having to analyze other languages than English, as most do not support this. Using this twitter address link for the four smartphones search queries I acquired four Excel datasets with Docteur Tweety each consisting of 50 tweets, ready to be analyzed by the tools.

4. RESULTS

4.1 Comparison of polarity tags

Figures 1-4 show the sentiment polarity tags analyzed by the various tools of the tweets regarding the Twitter data of the smartphone models. For better comparison I branded both the polarity tags P and P+ from Meaningcloud as positive and additionally N and N- as negative. In the same fashion are the tags positive, fairly positive and a little positive from Sentirate all branded as positive, and their negative counterparts as negative. Text2data was not able to analyze 13 tweets as it gave an unsupported language detection error. Somehow the tool recognized another language than English in the tweets. These 13 tweets are not included in the figures and explain why the Text2data numbers of Figures 2-4 do not add up to 50.

4.2 Comparison of sentiment strengths

As mentioned in the description of the tools, all tools except Umigon give a sentiment score to a text besides the polarity tag. This score ranges from -1.0 to +1.0 for all tools except Semantria. Their sentiment scores range from -2.0 to +2.0. To correctly compare the average sentiment scores I divided all Semantria scores by 2 causing them to be comparable with the scores from the other tools. Figures 5-8 show the average sentiment scores by the tools regarding the Twitter data of the smartphone models. The average strength of the scores provides additional insight combined with the distribution of the polarity tags to better interpret the results of the analyses.

\(^{15}\)https://twitter.com/search-advanced
4.3 Consistency of algorithms

To determine the consistency I have firstly calculated the differences of the polarity tags per tool per two datasets. For instance, the Semantria tool found 5 positive, 39 neutral and 6 negative sentiments in the iPhone 6 dataset whereas it found 3 positive, 46 neutral and 1 negative sentiments in the Samsung Galaxy S5 dataset. The difference can be easily found by subtracting the sum of the lowest numbers of the two datasets (3+39+1=43) from the total (50). When this is done for the various tools, we can calculate a mean and a median difference for these two datasets. These values are used in the succeeding step to calculate the consistency of the algorithms used by the tools. Table 1 shows the differences of the polarity tags between all dataset pairs for all the tools, calculated in the same fashion as depicted above.

Two notes should be added to this table before the consistency is calculated. Firstly, since Text2data could not analyze all 50 tweets in 3 of the 4 datasets, I have used the lowest total number of the two datasets compared for the total in the “formula” above. Secondly, I have abbreviated the smartphone models (iPhone 6 Plus to iP6P, Samsung Galaxy S5 to SG5, Nexus 6 to N6 and Sony Xperia Z3 to XZ3) for aesthetic purposes.

With the values acquired in the first step the deviations of the tool from the mean as well as the median can be calculated. This is done by subtracting the difference value of a tool from a dataset pair of the mean or median of this dataset pair. When done for all pair of datasets, the average deviation from both the mean and the median for each tool can be acquired. The result of this can be found in Figure 9.

5. DISCUSSION

5.1 Interpretation of results

With the results acquired from the various sentiment analyses statements about the concurrent validity and consistency of the tools can be made. With all but one tool providing their operational method, I can explain why some phenomena occur.

The first notable observation is the performance of Text2data. While the four other tools analyze mostly neutral sentiments, in all the datasets, this tool analyzes either a majority of positive or negative sentiments. Additionally, when the sentiment strengths are observed, it can be noted that Text2data seems to analyze more extreme sentiments when compared to the others. The other tools seem more in agreement about the sentiments, although some differences are notable. Semantria tends to classify almost all tweets as neutral. While Meaningcloud and Sentirate also classify the majority as neutral, a tendency is observable that they classify more tweets as positive. Umigon does the opposite, most
are analyzed as neutral, followed by negative. Looking at sentiment strengths the same observation can be made. When Meaningcloud and Sentirate are compared, Meaningcloud analyzes stronger positive sentiments to some extent.

While it is observable that only Text2data analyzes very different sentiments when compared to the other tools, this does not state that the tool is not accurate. The fact being that without knowing the nature of the tweets analyzed this statement cannot be made. After observing the datasets myself I concluded that most of the tweets indeed appear neutral, as they are objective advertisements or questions. Therefore I state that Text2data not only clearly differs from the other four tools used in this study, but also that their analysis seems inaccurate. In terms of test validity, this means that it appears that Text2data does not have content validity. As the other tools do not differ too much from each other and the majority of their sentiments analyzed are neutral, it seems they are quite accurate in general. However, since the exact sentimental values of the datasets are unknown, more than this cannot be said about their accuracy. This limitation is further discussed in Section 5.2.

Looking at the consistency of the various tools, which concerns the second research question, the same trend can be observed as with the concurrent validity, only in less extreme fashion. Text2data seems to be less consistent, mainly caused by the switch from analyzing mainly negative sentiments in the iPhone 6 Plus dataset to mainly positive in the others. Meaningcloud appears to have the most consistency compared to the others, although the difference is minimal. The other tools seem to have approximately equal consistency.

With all tools except Sentirate stating how their analyzing procedure works, some of the observations can be explained. Text2data combines four different approaches which can cause the tool to analyze too much sentiment. Since Text2data does not provide more information, it cannot be determined if this is either caused by one of the approaches, a wrong implementation, the combining procedure or another factor. The other three tools have in common that the text is decomposed into smaller parts, which are all given a sentiment score. These scores are then combined to determine the final document score. This is the whole procedure for Semantria. Umigon additionally looks at smileys and hashtags where Meaningcloud looks to firstly discriminate opinions from facts. Whether the minor differences between these three tools are caused by the algorithm they use in the procedure or the additional steps from Umigon and Meaningcloud cannot be determined. However, it seems that the operation of decomposing the text provides a concurrently valid and consistent sentiment analysis. This is not surprising as these three tools state to generally use the same procedure. The fact that these tools are in agreement proves that their procedure results in the same analysis in general, as there are minor deviations. Since Sentirate has not stated their procedure it can be the case that they either have a similar procedure or a different procedure that still results in approximately similar sentiment analyses.

5.2 Limitations

As mentioned in the previous section, the exact sentiment value of the used datasets is unknown. This fact limits this study as I can only conclude that all tools except Text2data seem to come close to being accurate. Future research in comparing sentiment analysis tools could make use of a dataset that is already sentimentally classified. This could be done by a set of linguistic experts. If this hand classified dataset is acquired, the values of these classifications can serve as golden standard in the concurrent validity comparison. This hand classified dataset can also be checked for internal consistency. Consequently, more precise statements can be made about the tools’ accuracy.

Furthermore, this study was limited by the use of trials, especially the ones from Text2data and the data gathering tool Docteur Tweety. This resulted in a small database to be analyzed. If more data was to be analyzed, the results could be made more credible.

6. CONCLUSION

This paper has set out to show the concurrent validity as well as the consistency of various sentiment analysis tools. With the use of the Docteur Tweety application Twitter data has been exported to .csv files to be analyzed by the tools. Five tools were found capable to analyze these external data sets: (1) Text2data, (2) Semantria, (3) Meaningcloud, (4) Sentirate, (5) Umigon. Results show that Text2data analyzes notably stronger sentiment than the other tools, which show minor differences when compared. Also Text2data seems the least consistent with their results when compared to the other tools. The Twitter data contained mostly objective advertisements and questions, which lead to the conclusion that Semantria, Meaningcloud, Sentirate and Umigon have more validity and reliability than Text2data.

This conclusion must be interpreted with some caution, as there are minor variations present in the sentiment ratios acquired by the tools when compared. Despite this, users of these tools now have (more) evidence that the sentiment tool they use, provides general accurate results as advertised. Perhaps this can better be phrased as there is no evidence found that the sentiment tool they use appears inaccurate, as is the case with Text2data. The limitations of this study prevent a more detailed conclusion. This paper also contributes to the theoretical research in the sentiment analysis field with two observations. Firstly, decomposing data in smaller parts to be analyzed for sentiment to later combine these sentiments for an overall sentiment seems to be a procedure that causes generally accurate results. Secondly, combining different sentiment analysis approaches seems to be a procedure that detects more sentiment than is actually present in a text and is therefore not accurate.

ACKNOWLEDGEMENTS

I would like to thank Dr. A.B.M.J. Wijnhoven for his guidance in this research. Furthermore, I wish to thank two student peer reviewers for their feedback.

7. REFERENCES


