University of Twente
School of Management & Governance

Bachelor Assignment

Vincent Poot
University of Twente, P.O. Box 217, 7500 AE Enschede,

E-mail: v.m.poot@student.utwente.nl
Telephone number: (+31)(0)6 24 33 79 97
Student number: s1010506

Supervisor: Dr. A.B.J.M. Wijnhoven
E-mail: fons.wijnhoven@utwente.nl
Telephone number: (+31)(0)5 34 89 38 53

Method for Evaluating Sentiment Analysis Tools

Abstract

Sentiment analysis tools are tools that extract subjective information from texts and process it into valuable information. A lot of these tools are available for individual users or small organizations (the laymen), who might not be familiar with these tools. This paper will start with a literature study on epistemological traditions and sentiment analysis for establishing criteria to evaluate the value of knowledge generated by sentiment analysis tools. Next a method for laymen to use these criteria for evaluating sentiment analysis tools is provided together with an evaluation of several available tools. This paper concludes that the method provided is useable for laymen in order to critically review a sentiment analysis tool and that developers of sentiment analysis tools need to take biases and their transparency into account when developing a sentiment analysis tools. Ultimately, recommendations for further research are given.

1. Introduction

With the rise of the world wide web more people can express their opinions online. Because of this an organization could for example gather and analyze information on what people say about their products online. According to a study by comScore and The Kelsey Group(comScore, 2007) more than 75% of the review users said that the review they read online had a significant impact on their purchase decision. They also stated that reviews made by other consumers made more impact than reviews by professionals. This shows the value consumers give to online product opinions from their peers and therefore shows what kind of value these opinions could have for organizations.

Because reviewing every opinion stated online impossible, methods and tools for acquiring these opinions and processing them into valuable information have been developed. These tools and services are called sentiment analysis tools. A sentiment analysis tool is a tool someone can use in order to find out what the opinion is on a certain topic. It extracts subjective information from texts in order to create valuable information for the user. The company can use publicly available information (like Twitter-tweets) to search for opinions on their product. A sentiment analysis tool could gather this information and classify and summarize found opinions and their sentiment. This information could then be used for the design for their next Smartphone. There are however a few challenges involved with the use of sentiment analysis. First of all there are the technical challenges. For example an opinion could be phrased as a double negative or a user also gives his opinion about something else in the same text, this could lead to a semantic classification problem. Then there is the problem of the external validity of sentiment analysis reports. Wijnhoven and Bloemen elaborate on this problem in their article (Wijnhoven & Bloemen, 2013). Three kinds of biases are discussed that may influence the author of the expressed opinion: demographics bias, manipulation of reviews and bias caused by events.

Sentiment analysis tools are used for different purposes, such as predicting the stock market, forecasting elections or gaining customer opinions on products. Some show to be quite accurate, such as predicting the stock market by analyzing daily Twitter feed and extracting the several mood dimensions and using this to predict the daily up and down changes in the closing values of the Dow Jones Industrial Average(Bollen, Mao, & Zeng, 2011). Others show to be less accurate, for example tools to predict elections. A recent study has shown that the published research methods for using Twitter data to predict elections are not better than chance (Metaxas, Mustafaraj, & Gayo-Avello, 2011).. This shows that further research in the limitations of these tools is necessary.

A lot of these tools are available for individual users and small organizations who don't necessarily know much about how the process of sentiment analysis works. These users are mainly interested in the result of this analysis rather than the technical details of the process. For these users it can be difficult to find the right tool for their purpose in order to achieve their end-goal in terms of knowledge generation. This study aims to provide a method for these users (the laymen) to evaluate the knowledge-value generated by sentiment analysis tools. The layman could for example be a small-business owner or anyone with an interest in

sentiment analysis but doesn't have much technical knowledge about these tools. Because a lot of these tools require technical and programming knowledge the laymen might not be able to evaluate the tool based on technical details. The key question thus is: "How can a layman critically review sentiment analysis tools and services in order to choose the right tool for his purpose?". To answer this question a literature study will be conducted on sentiment analysis tools and epistemological traditions. For the epistemological traditions Wijnhoven en Churchman's inquiry systems will be used. Each epistemological tradition has its own view on what is important in information. Based on these epistemological traditions criteria will be established in order to value information. These criteria can then be used to value the generated knowledge by sentiment analysis tools. Then a search and evaluation of existing tools will be given using the criteria established. Next the method will be tested for their usability by laymen. Finally, conclusions will be given with recommendations and suggestions for further research.

2. Literature study

In order to find literature on the use of sentiment analysis and sentiment analysis tools the following query, as suggested by (Wijnhoven & Bloemen, 2013) but slightly altered, has been used: *("Sentiment analysis") OR ("Opinion Mining") AND (("Social Media" OR "User generated content" OR reviews OR blog OR forum*)).* Using this query in Scopus (June 2013) gives 788 results. These results are then sorted by "Cited by" and from these I selected the first 30 papers that seemed relevant by title (see appendix I for an overview of these articles). I then proceeded to read the abstracts from these 30 papers and selected and read 6 papers (Cambria, Speer, Havasi, & Hussain, 2010; Chen & Tseng, 2011; Lee, 2008; Liu, 2010; Tsytsarau & Palpanas, 2012; Wu et al., 2010). The criteria for being "useful" in this research is whether the article is about the quality of the information of sentiment analysis and is written in non-technical terms (see table 1 for an oversight on the chosen articles and reason for its selection).

Author	Year	Title	Reason for selection
Tsytsarau, M., Palpanas, T.	2011	Survey on mining subjective data on the web	Gives an overview of current research and proposed methods
Chen, C.C., Tseng, Y.-D.	2011	Quality evaluation of product reviews using an information quality framework	Uses an information quality framework
Cambria, E., Speer, R., Havasi, C., Hussain, A.	2010	SenticNet: A publicly available semantic resource for opinion mining	Discusses a resource for opinion mining
Wu, Y., Wei, F., Liu, S., Au, N., Cui, W., Zhou, H., Qu, H.	2010	OpinionSeer: Interactive visualization of hotel customer feedback	Discusses a visualization system for sentiment analysis
Liu, B.	2010	Sentiment analysis: A multifaceted problem	Gives an overview of the field and challenges
Pang, B., Lee, L.	2008	Opinion mining and sentiment analysis	Gives a an overview on current research on sentiment analysis and opinion mining

Table 1: Chosen articles and reason for selection

The results are mainly papers that are very technical: only a few papers are within the scope of this research. When using ["Sentiment Analysis Tool"] as a query it only gives 12 results, with no articles relevant to this research. This suggests a gap in research on sentiment analysis tools and further research needs to be done. Because there is no research on sentiment analysis tools available for laymen, let alone on the quality of the generated knowledge by these sentiment analysis tools, this study will focus on the evaluation of the value of generated knowledge by sentiment analysis tools.

In order to evaluate the generated knowledge by sentiment analysis tools one has to look at the quality of the information it generates. To evaluate this knowledge, I will identify criteria based on 5 epistemological traditions. For the epistemological traditions I use (Wijnhoven, 2009), (Churchman, 1971), (Wijnhoven & Bloemen, 2013) and (Lacey, 1996) as an addition to Churchman.

3. Value of information

Wijnhoven identifies, on basis of Churchman, five epistemological traditions for the evaluation of knowledge: Lockean Empiricism, Leibnizian Rationalism, Kantian Idealism, Hegelian dialectics and the Singerian Pragmatism. These five traditions all have their own view on information and its value. These epistemological traditions will be used in order to establish criteria for the value of the generated knowledge from the sentiment analysis tools. These criteria can then be used by First I will shortly summarize every tradition and their ideas on information.

3.1. Lockean empiricism

Empiricism is a knowledge theory that assumes that knowledge can only (or mainly) be acquired by means of observation or experimentation. Lockean empiricism states that a human is born as a tabula rasa, a blank tablet, and assumes the human mind has no innate ideas and all knowledge is a posteriori. Because knowledge is the result of experience it is important that the measurements meet certain quality criteria: precision, accuracy, reliability and validity (Babbie, 2010). Precision refers to how precise a measurement is. For example: this book was written in the 1900's is less precise than when you say this book was written in 1902. A general rule of thumb is that more precise is better than less precise. Accuracy refers to how accurate the measure is, how well the observation reflects the real world. The book was written in The Netherlands is more precise than the book was written in Europe. If the book was written in Germany the latter statement, even though less precise, is more accurate than the first statement. Reliability of an observation means that the same data will be measured every time in a repeated observation (Babbie, 2010). Validity refers to whether a measurement reflects the concept that was intended to measure. For example if you are trying to measure nationalism you don't want to measure patriotism, even though they might be related, they are not the same. The demographics of the group you are mining opinions on is also important to take into account. The demographics may not represent the group you want to analyze and this could lead to a demographics bias and have an impact on the value of the generated knowledge. Therefore the sentiment analysis tool should take this bias in account in its analysis. These are all very important concepts in order to value information from an empirical perspective. Accuracy will not be used as criterion because this is not relevant when evaluating the generated knowledge of sentiment analysis tools because there is no way to check how accurate the measure actually reflects the real world. The validity criterion is used to see how well the measurement measures what was intended to measure. Because the data used by sentiment analysis tools is empirical it is important to use

3.2. Leibnizian Rationalism

Rationalism is a knowledge theory that assumes reason is the main source of knowledge or

"any view appealing to reason as a source of knowledge or justification"(Lacey, 1996). This view in its most extreme form beliefs that knowledge, contrary to empiricism, is obtained a priori. This means that there is no need for a sensory experience. Because of this assumption empiricism and rationalism seem to be opposites but actually don't necessarily have to be mutually exclusive as a philosopher can be both empiricist and rationalist (Lacey, 1996). The most important point is "that the creation of knowledge is not based on the development of consensus in a group of experts, but that any person with the proper kind of logical and reasoning capabilities may be able to discover knowledge and models of reality on his or her own"(Wijnhoven, 2009). This means that for information to be valuable, it must be the result of proper reasoning. Leibniz' principle of sufficient reason states that nothing is without reason and no effect is without a cause (Look, 2008). Conclusions should be drawn from proper argumentation and clear reasoning, so a requirement for information in order to be valuable is that the reasoning behind this generated knowledge should be detailed and understandable. A sentiment analysis tool should draw logical conclusions (logical consistency) from the obtained data and this reasoning must be detailed in order for it to be valuable.

3.3. Kantianism

Immanuel Kant explains his theory in "The Critique of Pure Reason". In this work he states that we should not only look at experience but at both experience and a priori concepts. This means that a sensory experience is influenced by an a priori concept. A way of looking at this is as if someone is wearing glasses, where the glasses represent the way they look at the world around them and other people can have different glasses that shape their sensory experiences and view on the outside world. Kant argued that one phenomenon could be seen from multiple perspectives and that not everybody looks at things the same. However, people will probably understand each other and can complement each other with the information obtained from their perspective. So in order for information generated by sentiment analysis tools to be valuable the tool should look at multiple categories. For example when looking at a car, the sentiment analysis tool can look at the price, the comfort and the speed of the car. This doesn't really give a clear image of what people think of this car, people might think this car is a bit overpriced, uncomfortable and not fast enough. They might, however, really appreciate other aspects of the car so that the complete image of the car is rather positive. So in order to give a relevant and complete view of what people think of this car the sentiment analysis needs to distinguish different categories. These categories should be relevant and complete in order for the generated knowledge to be valuable. It is also important to take event bias into account. An event could influence a person's "glasses" and therefore their sensory experience . Therefore another requirement is that the sentiment analysis tool takes this bias into account and tries to eliminate it.

3.4. Hegelian dialectics

Hegel developed a view on knowledge and created a dialectic system. This dialectic system is described by (Churchman, 1971) as a three-step process: thesis, anti-thesis and synthesis. Where the thesis is one view on a phenomenon and the anti-thesis a opposing view on the same matter. In order to solve this conflict a synthesis has to be found. There are however

always different views and interests in play. In order to find a synthesis it is key to eliminate possible sources of bias that might distort a critical analysis. One way to do this is to use triangulation. Triangulation is the use of multiple methods in order to increase the validity of the information (Wijnhoven, 2009). So in order to increase the value of information (the synthesis) sources of biases should be eliminated. (Wijnhoven & Bloemen, 2013) focuses on three kinds of biases found in opinion mining: demographics bias (issue of empirical nature), manipulation of reviews (Hegelian issue) and bias caused by events (Kantian issue). Manipulation of reviews is an Hegelian issue because it manipulates a critical analysis due to the illegitimate views it produces and therefore manipulates the synthesis. These biases are a threat to the value of the generated knowledge. Therefore the value of generated knowledge can be evaluated by looking at how biased this knowledge is and in what way these sentiment analysis tools try to eliminate these possible sources of bias.

3.5. Singerian Pragmatism

All of these aforementioned epistemologies share one characteristic: they all aim at finding the truth. The Singerian pragmatism integrates all of these epistemologies but thinks that finding the truth is only valuable when it improves the human condition (Churchman, 1971). It searches the knowledge that is useful for everyone, regardless of time and place. In this pragmatic view the focus lies on the means to an end; the generated knowledge only has value when it helps solving the issue at hand. This means that the knowledge generated by a sentiment analysis tool must be relevant for the issue at hand and display the generated knowledge in a way that is understandable for laymen.

4. Assessment method

In order to evaluate the generated knowledge by sentiment analysis tools this concept must be operationalized. After the operationalization I will provide a method for the layman to scale the presence of these criteria in the generated knowledge by the sentiment analysis tool. The presence of these criteria then indicate the value of the evaluated tool.

4.1. Criteria for the value of generation knowledge by sentiment analysis tools

Based on what the five epistemological traditions discussed in chapter 3 see as valuable information, I have identified 14 criteria that indicate the value of the generated knowledge by a sentiment analysis tool (see table 2 for a summary).

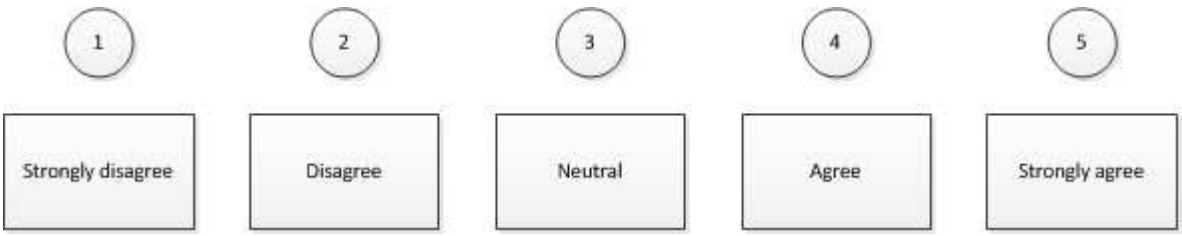


Figure 1: Likert scale

Each of these criteria will now briefly be discussed in chapter 4.2. In order to measure these criteria I use a Likert scale (see figure 1) in combination with a statement reflecting the

presence of this criterion. The layman can then score these criteria on that scale. Criteria with multiple statements reflecting the presence of that criterion will be rated the average score of those statements (rounded up to an even number).

	Epistemological Traditions				
Criteria	Lockean Empiricism	Leibnizian Rationalism	Kantianism	Hegelian dialectics	Singerian Pragmatism
Validity	x				
Reliablity	x				
Precision	x				
Causality		x	x		
Detail of reasoning		x			
Transparency of reasoning		x			
Understandability for laymen		x	x		x
Relevancy for problem solving					x
Usability of information					x
Transparency of dealing with biases	x (demographics bias)		x (event bias)	x (manipulation of reviews)	
Elimination of biases	x (demographics bias)		x (event bias)	x (manipulation of reviews)	
Categories	x		x		
-Relevancy		x	x		
-Completeness	x	x	x		

Table 2: The criteria indicating the value of generated knowledge by sentiment analysis tools

4.2. Operational definitions

For each criteria the operational definition will now be given. Each criterion will be measured by one or more statements reflecting the presence of that criterion. These statements are shown in table 3.

Validity
Validity refers to whether a measurement reflects the concept that it was intended to measure. Does the sentiment analysis tool give results about what you intended to measure or does it give results about something else?

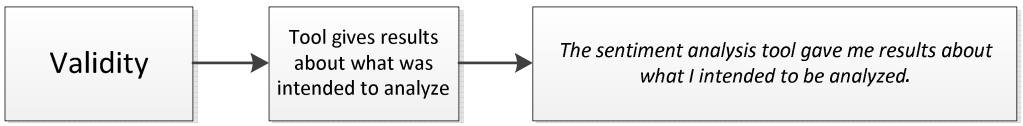


Figure 2: operationalization of validity

Reliability
Reliability of an observation means that the same data will be measured every time in a repeated observation. Does the sentiment analysis tool give the same results when the search is repeated?

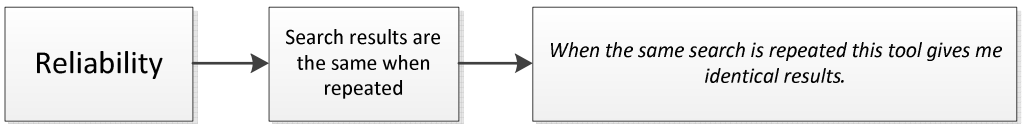


Figure 3: operationalization of reliability

Precision
Precision refers to how precise a measurement is. Are the results precise or are they vague? General rule of thumb here is that more precise is better than less precise.

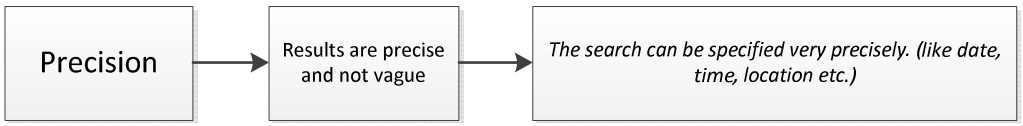


Figure 4: operationalization of precision

Relevancy for problem solving
Relevancy is about whether the generated knowledge is relevant for solving the problem at hand. Does the sentiment analysis tool provide you with information that is relevant to the problem you were trying to solve?

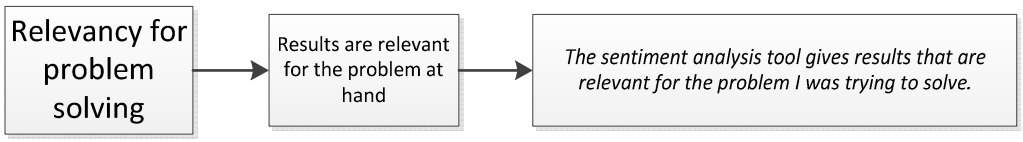


Figure 5: operationalization of relevancy for problem solving

Transparency of reasoning
Transparency of reasoning is about whether the sentiment analysis tool shows you how it came to its conclusions. Does the tool provide you with insight on how they get from the data to their conclusions or does it only give you the conclusions?

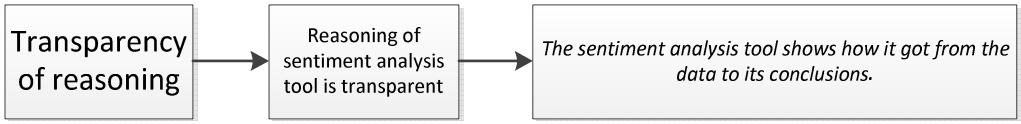


Figure 6: operationalization of transparency of reasoning

Logical consistency
Logical consistency refers to reasoning behind the generated knowledge. Is the reasoning logical? Or do the conclusions drawn in the analysis make no sense?

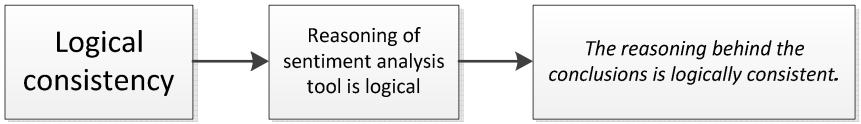


Figure 7: operationalization of causality

Detail of reasoning
Detail of reasoning refers to whether the reasoning from the sentiment analysis tool is detailed or not. It is important for the sentiment analysis tool to use detailed reasoning as reasoning that is too simple would give wrong conclusions

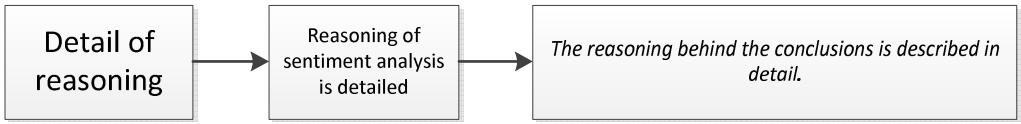


Figure 8: operationalization of detail of reasoning

Understandability for laymen
Understandability for laymen refers to whether the results are given in a way they are understandable for a layman. Are the results presented in non-technical terms and an understandable way?

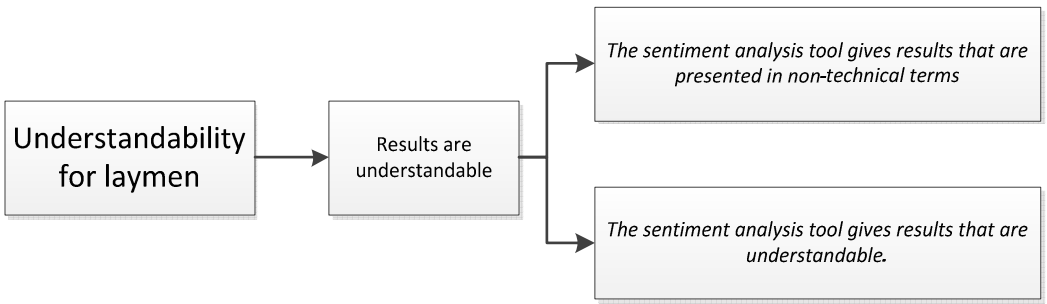


Figure 9: operationalization of understandability for laymen

Usability of information
Usability of information refers to whether the sentiment analysis tool produces usable information. Does the tool provide you with directly usable information?

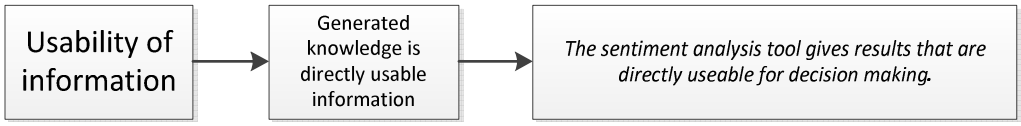


Figure 10: operationalization of usability of information

Categories
Categories refers to the number of categories the sentiment analysis tool provide you with. The sentiment analysis tool should look at enough categories.

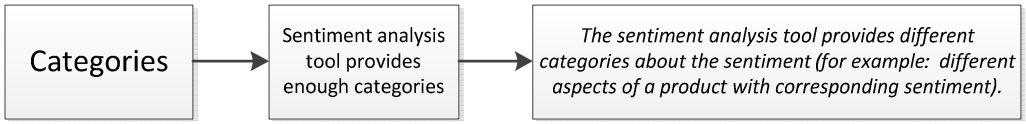


Figure 11: operationalization of categories

Relevancy of categories
Relevancy of categories refers to whether the categories the sentiment analysis tool provides you with are relevant

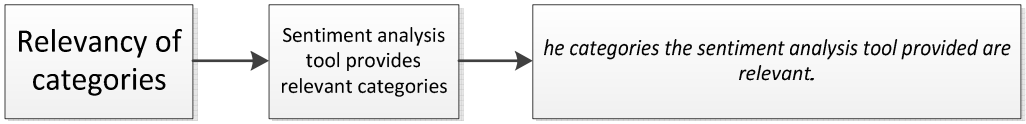


Figure 12: operationalization of relevancy of categories

Completeness of categories
Completeness of categories refers to whether the categories the sentiment analysis tool provides you with are complete

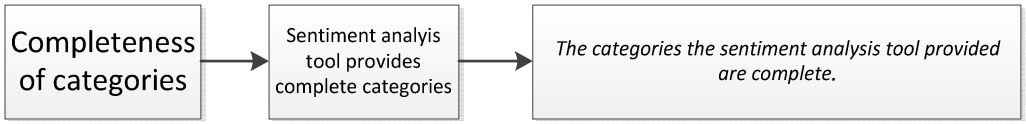


Figure 13: operationalization of completeness of categories

Transparency of dealing with biases
Transparency of dealing with biases refers to whether the sentiment analysis tool provides you with insight on how it deals with biases. Does the tool show you how it deals with biases?

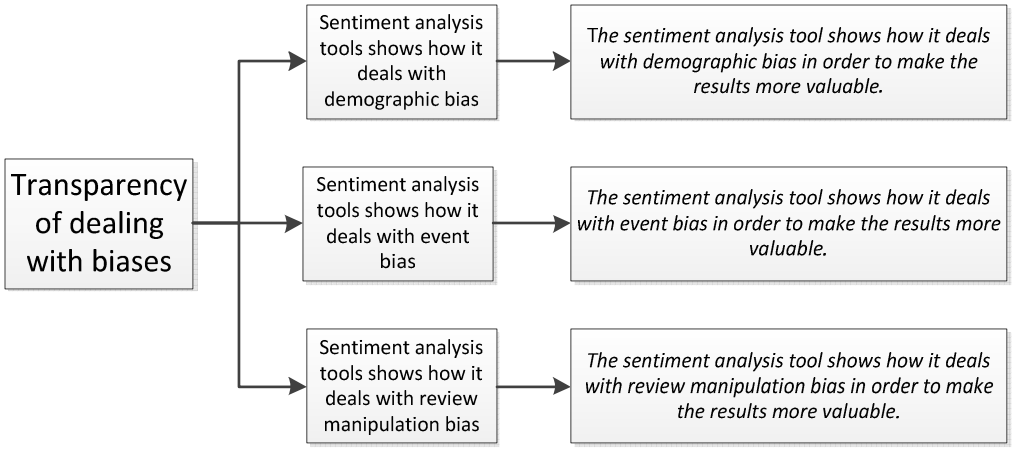


Figure 14: operationalization of transparency of dealing with biases

Elimination of bias

Elimination of bias refers to whether the sentiment analysis tool takes biases into account. Does it take demographic bias, event bias and manipulation bias into account when analyzing the data?

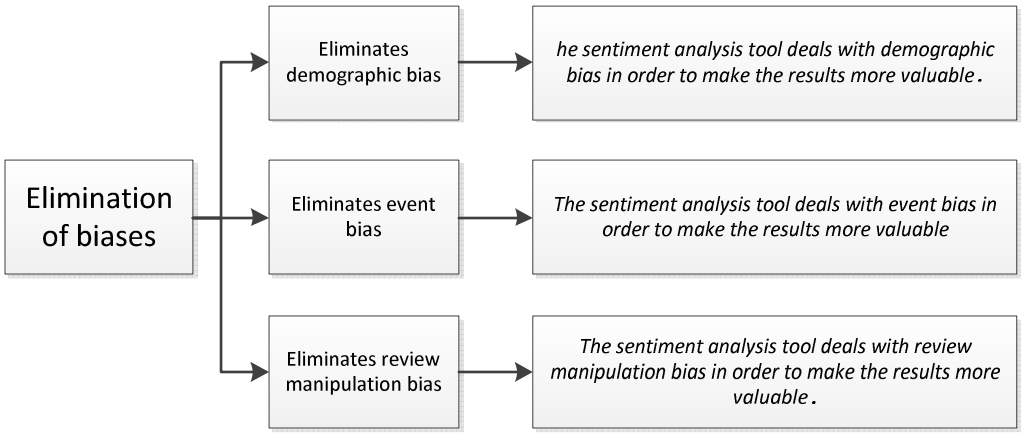


Figure 15: operationalization of elimination of biases

5. Evaluation of sentiment analysis tools

This chapter gives an evaluation of existing tools. First of all a search is done in for sentiment analysis tools that are usable for laymen. Then a brief description per tool will be given in combination with an evaluation using the method of chapter 4.

5.1. Search for sentiment analysis tools

Because no funds are available for these tools they are required to be either (1) available for free, (2) have an available trial version of their software or (3) have a demo available. See table 4 for an overview of the found tools.

Sentiment analysis tool	Free	Trial	Demo	URL
Topsy Pro Analytics		**x**		https://pro.topsy.com/
Veooz	**x**			http://www.veooz.com/
Opinion Crawl	**x**			http://www.opinioncrawl.com
Crimson Hexagon*			**x**	http://www.crimsonhexagon.com/
Sysomos**			**x**	http://www.sysomos.com/
General Sentiment**			**x**	http://www.generalsentiment.com/
Trackur		**x**		http://www.lymbix.com/
Sentiment140	**x**			http://www.sentiment140.com/
Socialmention	**x**			http://www.socialmention.com
Radian6			**x**	http://www.salesforcemarketingcloud.com/

Table 3: Sentiment analysis tools

**Demo is available according to website, but access was denied when I contacted them.*
***Demo is available according to website, but no access has been granted.*

5.2. Evaluation of sentiment analysis tools

For each evaluation the same query has been used with the same end-goal. I searched for the topic: "iPhone 5" in order to see what the sentiment on this topic is so I could find out what needs to improve with the next iPhone in order to gain more customer satisfaction. This has been done for the purpose of rating the criteria. See table 5 for a overview of the tools and the criteria with their scores and see appendix IV for screenshots of the output.

Topsy Pro Analytics
Topsy Pro Analytics is a paid service that gathers and analyzes social web data. It analyzes hundreds of billions of tweets from Topsy's index. It gives information on trending topics, opinion leaders and provides sentiment per topic. You can enter a search query and specify this on date, location, language, sentiment and more. It is also possible to do a comparative search for example to compare different companies in the same industry. Topsy will identify the most important and influential posts and users.

Veooz
Veooz is a simple sentiment analysis tool. It uses Twitter, Facebook and news comments as sources and shows the sentiment for the specified topic for (up to) the last 90days. It also gives related news, photos and videos for the entered query. It allows you to filter the activity per source, opinion and gender. It is easy to use but does not give a lot of relevant information and statistics.

Opinion Crawl
Opinion Crawl is a simple and easy to use sentiment analysis tool. It is available for free and allows to search the web for a topic and gives the sentiment for that topic. It also provides recent news and key concepts on the specified query. It does not provide any over-time statistics and therefore the information is not very relevant. It also does not show what the tool exactly analyzes in order to calculate the sentiment.

Trackur
Trackur is a paid service that monitors social media. It uses news, blogs, Facebook, Twitter, Google+ and other social media sites as sources. It allows you to enter a search query and gives you the results about recent posts about that topic. It shows the sentiment, source, influence and date per result. You can specify your query by country and source. It shows sentiment trends for up to 7 days and identifies how much influence a post has.

Sentiment140
Sentiment140 is a free and simple sentiment analysis tool. It uses Twitter in order to obtain the sentiment for the specified topic. It is possible to specify for English or Spanish tweets. It shows the overall sentiment by percent and shows recent tweets about the topic. It only gives the sentiment on tweets for the past hour and does not give any over-time statistics and does not identify influential users or posts.

Socialmention
Social mention is a free sentiment analysis tool. It does real-time social media searches and analysis. It searches through the web for mentions on the specified topic. It shows recent mentions (up to a month), the sentiment, top users and sources. It is also possible to specify the search per source (blogs, microblogs, comments, events, images, news, video, audio, Q&A and networks) and date.

Radian6
Radian6 is a paid service that monitors social media posts, blogs, news sites, discussion boards and video websites. It identifies and analyses conversations about your company, product or competitors. Radian6 provides the user with real-time information on topic, the sentiment and identifies key users. Multiple languages are supported and results can be filtered by time, location, source and media type. It allows you to see influential posts and allows you to respond to these posts. Multiple functions are available in order to make it easy to use and allows for multiple users to use the same account.

Criteria	Topsy	Veooz	Opinion crawl	Trackur	Sentiment140	Socialmention	Radian6
Empiricism							
-Validity	5	5	5	5	5	5	5
-Reliability	5	5	5	3	5	5	5
-Precision	4	2	2	3	2	3	5
-Elimination of biases (demographics bias)	1	1	1	1	1	1	2
-Transparency of dealing with bias (demographics bias)	1	1	1	1	1	1	2
Rationalism							
-Logical consistency	2	1	3	1	5	1	1
-Detail of reasoning	5	1	4	1	5	1	1
-Transparency of reasoning	1	1	4	1	1	1	2
Kantianism							
-Transparency of dealing with biases (event bias)	1	1	1	1	1	1	2
-Elimination of biases (event bias)	1	1	1	1	1	1	2
-Categories	1	1	1	1	1	1	1
-Relevancy of categories	1	1	1	1	1	1	1
-Completeness of categories	1	1	1	1	1	1	1
Hegelian dialectics							
-Transparency of dealing with biases (manipulation of reviews)	1	1	1	1	1	1	2
Elimination of biases (manipulation of reviews)	1	1	1	1	1	1	2
Singerian pragmatism							
-Understandability for laymen	5	5	5	4	5	4	5
-Relevancy for problem solving	5	3	2	4	2	2	5
-Usability of information	4	2	3	4	2	4	5

Table 4: *overview of evaluated tools*

6. Usability for laymen and inter-rater reliability

In this chapter an evaluation for the usability of the method for laymen will be given. In order to evaluate the usability of the assessment method for laymen a weighted kappa κ with linear weights will be used to test the inter-rater reliability. Three laymen will score the sentiment analysis tools Sentiment140 and Topsy Pro Analytics. Sentiment140 is a free tool and Topsy is a paid tool. The scores will be compared to my own scores (user-expert rater) for the same sentiment analysis tool. A user-expert rater is a rater that has used multiple sentiment analysis tools and has some experience in using these tools, in this case myself. All four raters will use the same search query and end-goal (the same as described in chapter 5.2). The expert rating and rater 2 will be tested for inter-rater reliability, rater 2 and 3 will be tested for inter-rater reliability and rater 1 and 3 will also be tested for inter-rater reliability. This way the usability of the tool for laymen will be tested together with the inter-rater reliability of the scale (see figure 15). The scores per rater and screenshots of the inter-rater reliability tests can respectively be found in appendix II and III.

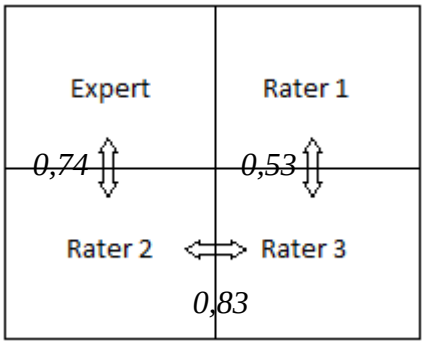


Figure 15: IRR Sentiment140

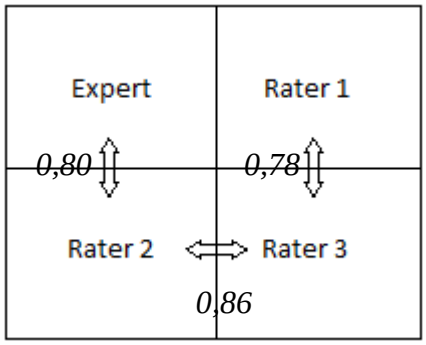


Figure 16: IRR Topsy Pro

As (Landis & Koch, 1977) proposed, the following labels will be used to interpret the strength of kappa: <0.00 is poor agreement, 0.00-0.20 is a slight agreement, 0.21-0.40 is a fair agreement, 0.41-0.60 is a moderate agreement, 0.61-0.80 is a substantial agreement and 0.81-1.00 is an almost perfect agreement. A requirement for this method is that it is usable by laymen. This means that the expert rater and the laymen rater should have an inter-rater reliability of at least κ=0.61.

For Sentiment140 expert and rater 2 have an inter-rater reliability of κ=0.74, which indicates substantial agreement between the raters. This shows that the expert rater and the layman rater have a substantial agreement and this means the scale is usable for laymen. Rater 1 and 3 have an inter-rater reliability of κ=0.53, which is a moderate agreement. Rater 2 and 3 have an inter-rater reliability of κ=0.83, which is an almost perfect agreement. Both of these are acceptable and indicate that the scale is appropriate for measuring the criteria.

For Topsy Pro Analytics expert and rater 2 have an inter-rater reliability of κ =0,84. Which indicates an almost perfect agreement. Rater 1 and 3 have an inter-rater reliability of κ=0,78, which is a substantial agreement and rater 2 and 3 have a inter-rater reliability of κ=0,86.

The difference between the moderate IRR between rater 1 and 3 for Sentiment140 and the substantial IRR between rater 1 and 3 for Topsy Pro analytics can be explained by how these raters have rated both tools. If you look at the criteria individually you can see they both rated these tools quite similar but there are really small differences in rating per criterion.

7. Conclusions

The main research question of this paper is: "How can a layman critically review sentiment analysis tools and services in order to choose the right tool for their purpose?". In this paper a method for laymen in order to evaluate sentiment analysis tools is presented. This method consists of criteria that indicate the value of information which is made score-able for a layman in order to evaluate the value of the generated knowledge by sentiment analysis tools. Without much technical knowledge about sentiment analysis tools a layman can still critically review the sentiment analysis tool and evaluate a sentiment analysis tool in the same way a user-expert can. By looking at these criteria, which are based on epistemological traditions, he can decide whether the generated knowledge by the sentiment analysis tool is valuable or not. He can then decide which sentiment analysis tool is right for his end-goal. Furthermore the evaluation and description of the sentiment analysis tools given in this research will allow more people to find a sentiment analysis tool that is useable for laymen and will enable more people to use the right sentiment analysis tool. In general the paid tools score much better on all criteria except for transparency. This could be explained by the fact that the paid services want to keep their process a secret in order to protect their product.

7.1. Recommendations

This research shows what criteria are important for the value of information. I recommend that developers of sentiment analysis tools look at these criteria order to improve the value of the generated knowledge of their sentiment analysis tool. It gives insight in what is important to users of the sentiment analysis tool. I recommend developers look at the criteria based on rationalism because developers are not very transparent in how their sentiment analysis tools get their results.

Furthermore I recommend that more research should be done on the usability of sentiment analysis tools for laymen. While there are quite some free sentiment analysis tools available which are very advanced and give better and more detailed results than free online sentiment analysis tools, they are not usable by laymen. They require substantial technical knowledge or programming experience for someone to use them. In order to enable more people to use sentiment analysis tools more research needs to be done on what makes a sentiment analysis tool usable for laymen so that designers of these tools can take that into account. Furthermore almost all sentiment analysis tools score badly on dealing with the biases, so this should also be addressed. Some tools might deal with these biases but don't show whether they do it or how they do it.

7.2 Discussion

Most of the evaluated tools score good on the empirical criteria (except for dealing with the demographical bias and its transparency). None of the tools scored good on the Hegelian and Kantian criteria which might indicate current tools focus on obtaining data and doing good analysis but lack focus on dealing with biases and analyzing the sentiment in categories. This means these criteria might not be useful for evaluating sentiment analysis tools at this moment but this does not mean these are not important criteria for the value of sentiment analysis tools. These criteria might be useful in the future when sentiment analysis tools are more developed and widely available for laymen.

References

Babbie, E. (2010). *The practice of social research*: Wadsworth, Cengage Learning.

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

Cambria, E., Speer, R., Havasi, C., & Hussain, A. (2010). *Senticnet: A publicly available semantic resource for opinion mining*.

Chen, C.C., & Tseng, Y.-D. (2011). Quality evaluation of product reviews using an information quality framework. *Decision Support Systems, 50*(4), 755-768.

Churchman, C.W. (1971). *The design of inquiring systems: Basic concepts of systems and organization*. New York: Basic Books.

comScore. (2007). Online consumer-generated reviews have significant impact on offline purchase behavior. Retrieved 25 July, 2013, from: http://www.comscore.com/Insights/Press_Releases/2007/11/Online_Consumer_Reviews_Impact_Offline_Purchasing_Behavior

Lacey, A.R. (1996). A dictionary of philosophy (3rd edition ed., pp. 286). London, UK: Routledge.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Lee, B.P.L. (2008). Opinion mining and sentiment analysis *Foundations and Trends in Information Retrieval* (Vol. 2, pp. 1-135).

Liu, B. (2010). Sentiment analysis: A multifaceted problem. *IEEE Intelligent Systems, 25*(3), 76-80.

Look, B.C. (2008). Gottfried wilhelm leibniz *Stanford Encyclopedia of Philosophy (Spring 2008 Edition)*.

Metaxas, P.T., Mustafaraj, E., & Gayo-Avello, D. (2011). How (not) to predict elections. 165-171.

Tsytsarau, M., & Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery, 24*(3), 478-514.

Wijnhoven, A.B.J.M. (2009). *Information management: An informing approach*: New York: Routledge.

Wijnhoven, A.B.J.M., & Bloemen, O. (2013). *External validity of sentiment mining reports: Challenges of demographics biases, events, and manipulation.* University of Twente.

Wu, Y., Wei, F., Liu, S., Au, N., Cui, W., Zhou, H., & Qu, H. (2010). Opinionseer: Interactive visualization of hotel customer feedback. *IEEE Transactions on Visualization and Computer Graphics, 16*(6), 1109-1118.

Appendix I

Scopus Documents: first 30 selected papers
==========
Thelwall M., Buckley K., Paltoglou G.
Sentiment strength detection for the social web
(2012) Journal of the American Society for Information Science and Technology, . Cited 11 times.

Tsytsarau M., Palpanas T.
Survey on mining subjective data on the web
(2011) Data Mining and Knowledge Discovery, . Article in Press. Cited 5 times.

Nikolay A., Anindya G., Panagiotis G.I.
Deriving the pricing power of product features by mining consumer reviews
(2011) Management Science, . Cited 14 times.

Ghose A., Ipeirotis P.G.
Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics
(2011) IEEE Transactions on Knowledge and Data Engineering, . Cited 23 times.

Xu K., Liao S.S., Li J., Song Y.
Mining comparative opinions from customer reviews for Competitive Intelligence
(2011) Decision Support Systems, . Cited 15 times.

Bai X.
Predicting consumer sentiments from online text
(2011) Decision Support Systems, . Cited 21 times.

Chen C.C., Tseng Y.-D.
Quality evaluation of product reviews using an information quality framework
(2011) Decision Support Systems, . Cited 11 times.

Cambria E., Speer R., Havasi C., Hussain A.
SenticNet: A publicly available semantic resource for opinion mining
(2010) AAAI Fall Symposium - Technical Report, . Cited 8 times.

Cambria E., Hussain A., Havasi C., Eckl C.
SenticSpace: Visualizing opinions and sentiments in a multi-dimensional vector space
(2010) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), . Cited 7 times.

Wu Y., Wei F., Liu S., Au N., Cui W., Zhou H., Qu H.
OpinionSeer: Interactive visualization of hotel customer feedback
(2010) IEEE Transactions on Visualization and Computer Graphics, . Cited 9 times.

Chen H., Zimbra D.
AI and opinion mining

(2010) IEEE Intelligent Systems, . Cited 17 times.

Liu B.
Sentiment analysis: A multifaceted problem
(2010) IEEE Intelligent Systems, . Cited 13 times.

Li N., Wu D.D.
Using text mining and sentiment analysis for online forums hotspot detection and forecast
(2010) Decision Support Systems, . Cited 39 times.

Danescu-Niculescu-Mizil C., Kossinets G., Kleinberg J., Lee L.
How opinions are received by online communities: A case study on Amazon.com helpfulness votes
(2009) WWW'09 - Proceedings of the 18th International World Wide Web Conference, . Cited 39 times.

Lin C., He Y.
Joint sentiment/topic model for sentiment analysis
(2009) International Conference on Information and Knowledge Management, Proceedings, . Cited 49 times.

Jin W., Ho H.H., Srihari R.K.
OpinionMiner: A novel machine learning system for web opinion mining and extraction
(2009) Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, . Cited 14 times.

Prabowo R., Thelwall M.
Sentiment analysis: A combined approach
(2009) Journal of Informetrics, . Cited 49 times.

Zhan J., Loh H.T., Liu Y.
Gather customer concerns from online product reviews - A text summarization approach
(2009) Expert Systems with Applications, . Cited 29 times.

Akehurst G.
User generated content: The use of blogs for tourism organisations and tourism consumers
(2009) Service Business, . Cited 32 times.

Zhang M., Ye X.
A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval
(2008) ACM SIGIR 2008 - 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Proceedings, . Cited 28 times.

Ganapathibhotla M., Liu B.
Mining opinions in comparative sentences
(2008) Coling 2008 - 22nd International Conference on Computational Linguistics, Proceedings of the Conference, . Cited 16 times.

Pang B., Lee L.
Opinion mining and sentiment analysis
(2008) Foundations and Trends in Information Retrieval, . Cited 683 times.

Abbasi A., Chen H., Salem A.
Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums
(2008) ACM Transactions on Information Systems, . Cited 108 times.

Ding X., Liu B., Yu P.S.
A holistic lexicon-based approach to opinion mining
(2008) WSDM'08 - Proceedings of the 2008 International Conference on Web Search and Data Mining, . Cited 101 times.

Archak N., Ghose A., Ipeirotis P.G.
Show me the money!: Deriving the pricing power of product features by mining consumer reviews
(2007) Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, . Cited 36 times.

Conrad J.G., Schilder F.
Opinion mining in legal blogs
(2007) Proceedings of the International Conference on Artificial Intelligence and Law, . Cited 21 times.

Kobayashi N., Inui K., Matsumoto Y.
Opinion mining from web documents: Extraction and structurization
(2007) Transactions of the Japanese Society for Artificial Intelligence, . Cited 8 times.

Yi J., Niblack W.
Sentiment mining in WebFountain
(2005) Proceedings - International Conference on Data Engineering, . Cited 25 times.

Whitelaw C., Garg N., Argamon S.
Using appraisal groups for sentiment analysis
(2005) International Conference on Information and Knowledge Management, Proceedings, . Cited 78 times.

Yi J., Nasukawa T., Bunescu R., Niblack W.
Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques
(2003) Proceedings - IEEE International Conference on Data Mining, ICDM, . Cited 99 times.

Appendix II

Criterion	*Score*	*Criterion*	*Score*
Validity	*5*	Relevancy for problem solving	*2*
Reliablity	*5*	Usability of information	*2*
Precision	*2*	Transparency of dealing with biases	*1*
Logical consistency	*5*	Elimination of biases	*1*
Detail of reasoning	*5*	Categories	*1*
Transparency of reasoning	*5*	Relevancy of categories	*1*
Understandability for laymen	*5*	Completeness of categories	*1*

Expert rater evaluation of Sentiment140

Criterion	*Score*	*Criterion*	*Score*
Validity	*4*	Relevancy for problem solving	*3*
Reliablity	*5*	Usability of information	*1*
Precision	*3*	Transparency of dealing with biases	*1*
Logical consistency	*4*	Elimination of biases	*1*
Detail of reasoning	*3*	Categories	*1*
Transparency of reasoning	*4*	Relevancy of categories	*1*
Understandability for laymen	*5*	Completeness of categories	*1*

Layman rater 1 evaluation of Sentiment140

Criterion	*Score*	*Criterion*	*Score*
Validity	*5*	Relevancy for problem solving	*1*
Reliablity	*4*	Usability of information	*3*
Precision	*1*	Transparency of dealing with biases	*1*
Logical consistency	*4*	Elimination of biases	*1*
Detail of reasoning	*4*	Categories	*1*
Transparency of reasoning	*4*	Relevancy of categories	*1*
Understandability for laymen	*5*	Completeness of categories	*1*

Layman rater 2 evaluation of Sentiment140

Criterion	*Score*	*Criterion*	*Score*
Validity	*5*	Relevancy for problem solving	*1*
Reliablity	*4*	Usability of information	*3*
Precision	*1*	Transparency of dealing with biases	*2*
Logical consistency	*5*	Elimination of biases	*1*
Detail of reasoning	*5*	Categories	*1*
Transparency of reasoning	*4*	Relevancy of categories	*1*
Understandability for laymen	*5*	Completeness of categories	*1*

Layman rater 3 evaluation of Sentiment140

Criterion	*Score*	*Criterion*	*Score*
Validity	*5*	Relevancy for problem solving	*5*
Reliablity	*5*	Usability of information	*4*
Precision	*4*	Transparency of dealing with biases	*1*
Logical consistency	*2*	Elimination of biases	*1*
Detail of reasoning	*5*	Categories	*1*
Transparency of reasoning	*1*	Relevancy of categories	*1*
Understandability for laymen	*5*	Completeness of categories	*1*

Expert rater evaluation of Topsy Pro Analytics

Criterion	*Score*	*Criterion*	*Score*
Validity	*5*	Relevancy for problem solving	*5*
Reliablity	*4*	Usability of information	*5*
Precision	*4*	Transparency of dealing with biases	*1*
Logical consistency	*1*	Elimination of biases	*1*
Detail of reasoning	*5*	Categories	*1*
Transparency of reasoning	*1*	Relevancy of categories	*1*
Understandability for laymen	*5*	Completeness of categories	*1*

Layman rater 1 evaluation of Topsy Pro Analytics

Criterion	*Score*	*Criterion*	*Score*
Validity	*4*	Relevancy for problem solving	*5*
Reliablity	*5*	Usability of information	*5*
Precision	*5*	Transparency of dealing with biases	*1*
Logical consistency	*1*	Elimination of biases	*1*
Detail of reasoning	*5*	Categories	*1*
Transparency of reasoning	*1*	Relevancy of categories	*1*
Understandability for laymen	*5*	Completeness of categories	*1*

Layman rater 2 evaluation of Topsy Pro Analytics

Criterion	*Score*	*Criterion*	*Score*
Validity	*5*	Relevancy for problem solving	*4*
Reliablity	*5*	Usability of information	*5*
Precision	*5*	Transparency of dealing with biases	*1*
Logical consistency	*1*	Elimination of biases	*1*
Detail of reasoning	*5*	Categories	*1*
Transparency of reasoning	*1*	Relevancy of categories	*1*
Understandability for laymen	*5*	Completeness of categories	*1*

Layman rater 3 evaluation of Topsy Pro Analytics

Appendix III

Inter-rater agreement (kappa)

Observer A	Rater_1
Observer B	Rater_3

Observer B	Observer A 1	2	3	4	5	
1	4	0	2	0	0	6 (42,9%)
2	1	0	0	0	0	1 (7,1%)
3	1	0	0	0	0	1 (7,1%)
4	0	0	0	1	1	2 (14,3%)
5	0	0	1	2	1	4 (28,6%)
	6 (42,9%)	0 (0,0%)	3 (21,4%)	3 (21,4%)	2 (14,3%)	14

Weighted Kappa[a]	0,528
Standard error	0,130
95% CI	0,273 to 0,783

[a] Linear weights

Frequency chart

Sentiment 140 rater 1 and 3

Inter-rater agreement (kappa)

Observer A	Rater_2 Rater 2
Observer B	Rater_3 Rater 3

Observer B	Observer A 1	2	3	4	5	
1	6	0	0	0	0	6 (42,9%)
2	1	0	0	0	0	1 (7,1%)
3	0	0	1	0	0	1 (7,1%)
4	0	0	0	2	0	2 (14,3%)
5	0	0	0	2	2	4 (28,6%)
	7 (50,0%)	0 (0,0%)	1 (7,1%)	4 (28,6%)	2 (14,3%)	14

Weighted Kappa[a]	0,883
Standard error	0,058
95% CI	0,770 to 0,997

[a] Linear weights

Frequency chart

Sentiment 140 rater 2 and 3

Inter-rater agreement (kappa)

Observer A	Expert
Observer B	Rater_2 Rater 2

Observer B	Observer A 1	2	3	4	5	
1	5	2	0	0	0	7 (50,0%)
2	0	0	0	0	0	0 (0,0%)
3	0	1	0	0	0	1 (7,1%)
4	0	0	0	0	4	4 (28,6%)
5	0	0	0	0	2	2 (14,3%)
	5 (35,7%)	3 (21,4%)	0 (0,0%)	0 (0,0%)	6 (42,9%)	14

Weighted Kappa[a]	0,742
Standard error	0,068
95% CI	0,610 to 0,875

[a] Linear weights

Frequency chart

Sentiment 140 expert rater and rater 2

Inter-rater agreement (kappa)

Observer A	Rater_1
Observer B	Rater_3 Rater 3

Observer B	Observer A 1	4	5	
1	7	0	0	7 (50,0%)
4	0	0	1	1 (7,1%)
5	0	2	4	6 (42,9%)
	7 (50,0%)	2 (14,3%)	5 (35,7%)	14

Weighted Kappa[a]	0,781
Standard error	0,110
95% CI	0,565 to 0,998

[a] Linear weights

Frequency chart

Topsy pro analytics rater 1 and 3

Inter-rater agreement (kappa)

Observer A	Rater_2
Observer B	Rater_3 Rater 3

Observer B	Observer A			
	1	4	5	
1	7	0	0	7 (50,0%)
4	0	0	1	1 (7,1%)
5	0	1	5	6 (42,9%)
	7 (50,0%)	1 (7,1%)	6 (42,9%)	14

Weighted Kappa[a]	0,856
Standard error	0,094
95% CI	0,671 to 1,000

[a] Linear weights

Frequency chart

Topsy pro analytics rater 2 and 3

Inter-rater agreement (kappa)

Observer A	Expert
Observer B	Rater_2 Rater 2

Observer B	Observer A				
	1	2	4	5	
1	6	1	0	0	7 (50,0%)
2	0	0	0	0	0 (0,0%)
4	0	0	0	1	1 (7,1%)
5	0	0	2	4	6 (42,9%)
	6 (42,9%)	1 (7,1%)	2 (14,3%)	5 (35,7%)	14

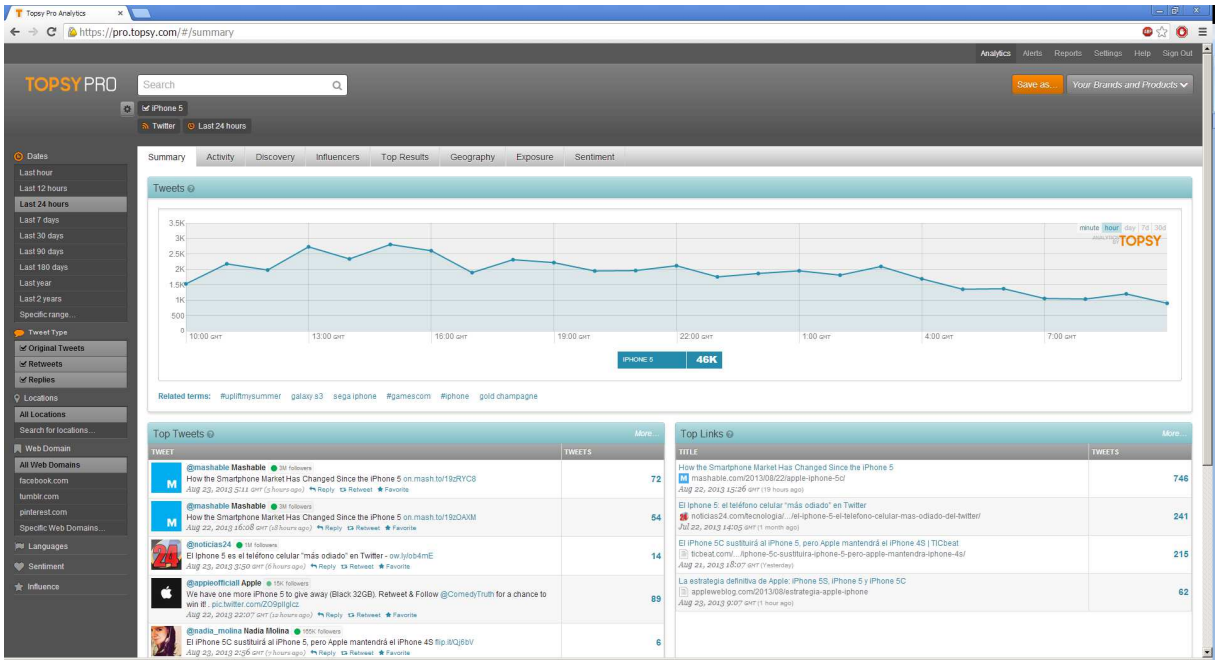
Weighted Kappa[a]	0,807
Standard error	0,082
95% CI	0,645 to 0,968

[a] Linear weights

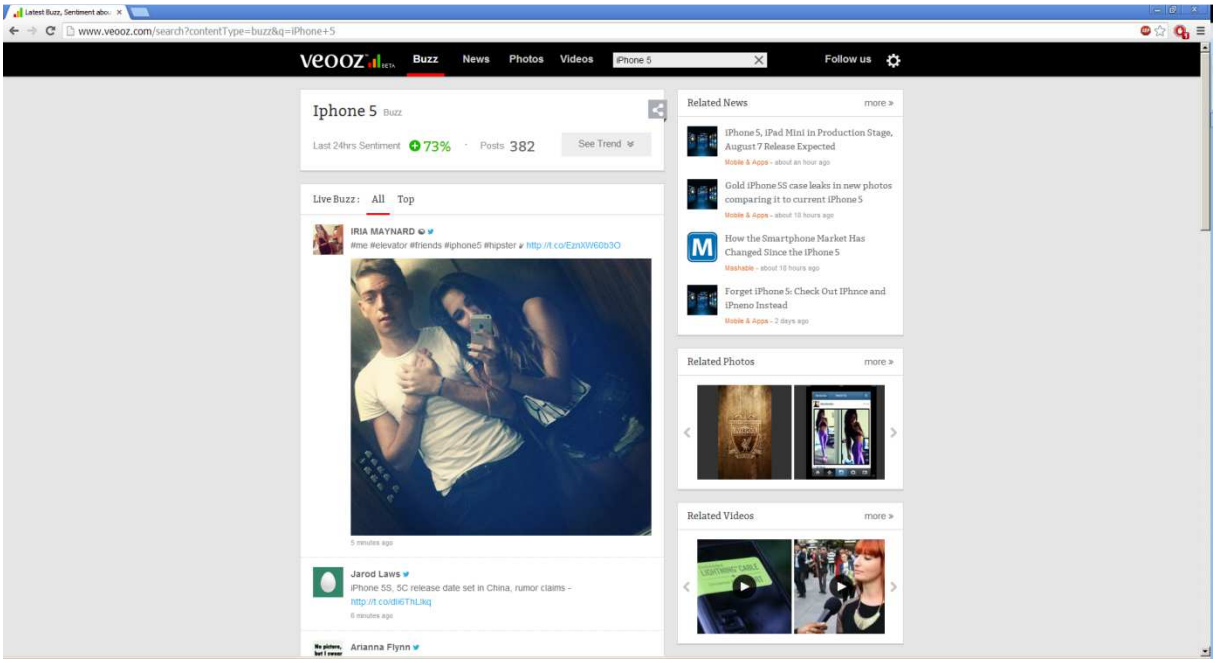
Frequency chart

Topsy pro analytics expert rater and rater 2

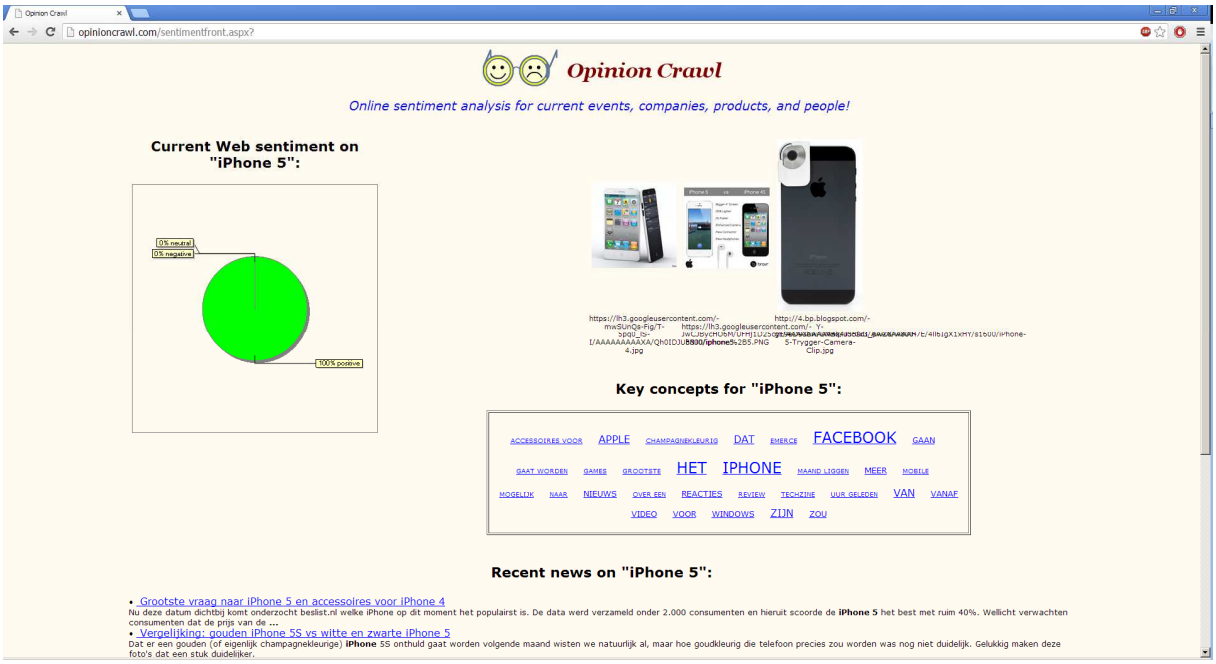
Appendix IV



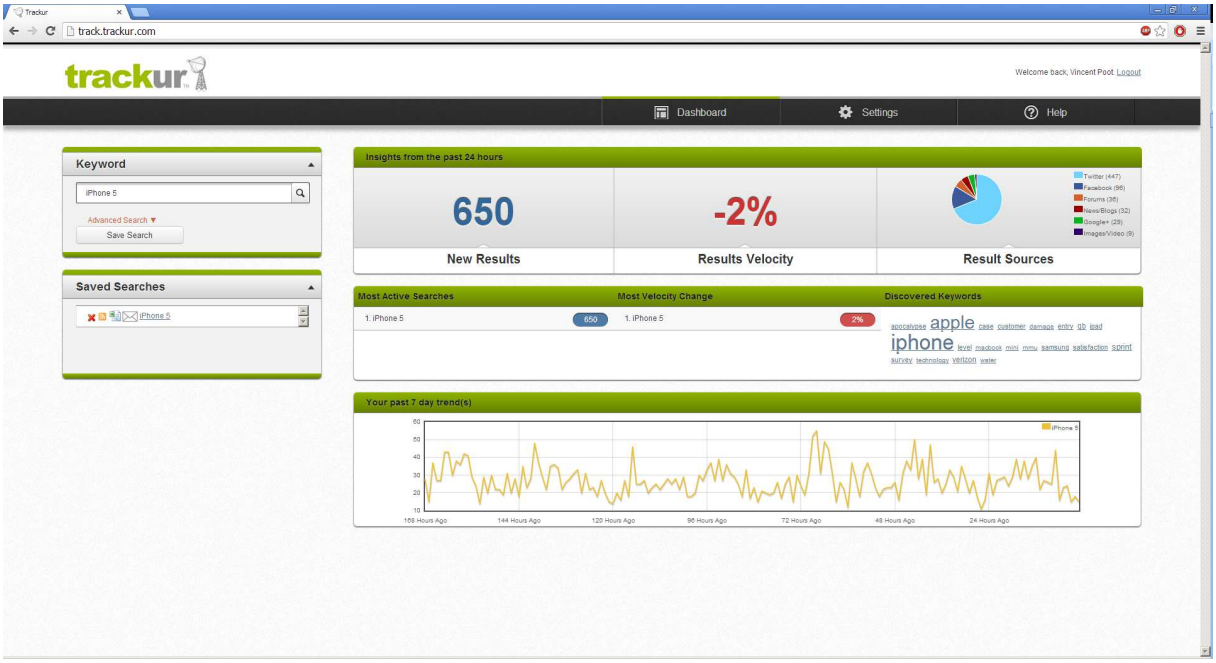
Topsy pro analytics



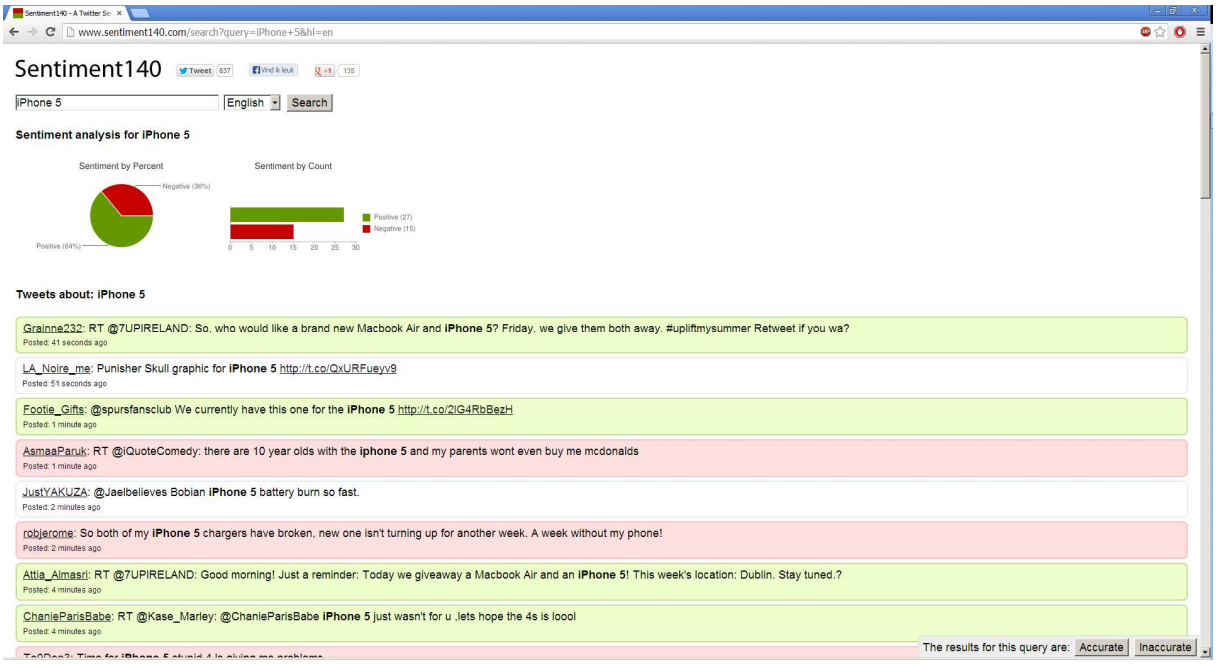
Veooz



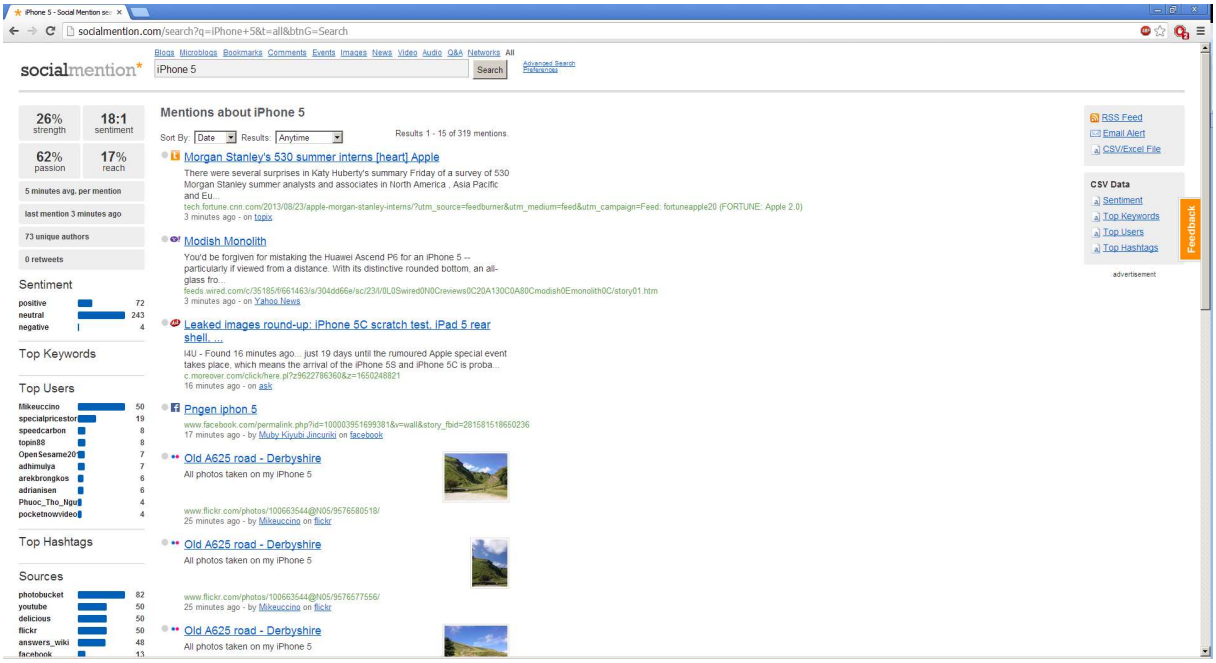
Opinion crawl



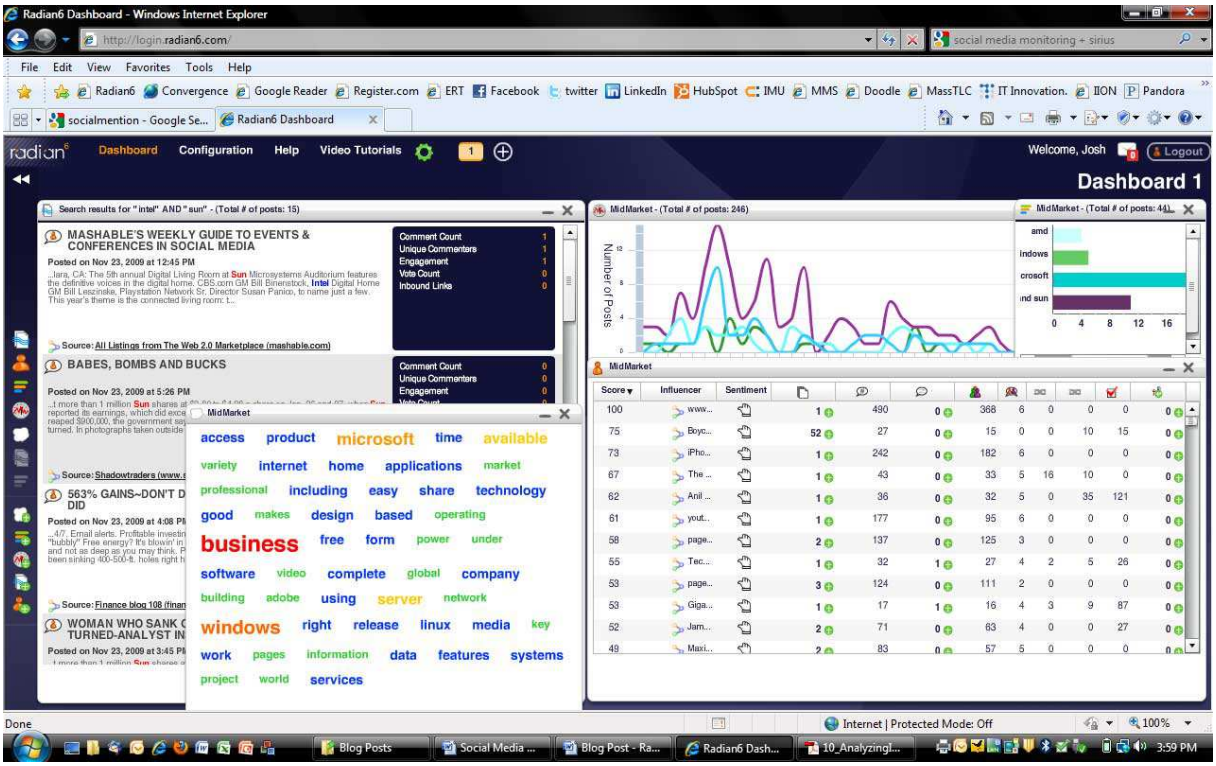
Trackur



Sentiment 140



Social mention



Radian6 (screenshot obtained from www.ombud.com at 22-8-2013, due to no access at this time)