Viability of sentiment analysis in business
Evaluating accuracy and the supporting NLP technologies

Peter Halling Hilborg
&
Emil Blædel Nygaard
1.0 Executive Summary

Natural language processing (NLP) has in later years seen more and more implementations in businesses, largely because of sentiment analysis and its applications to social data, such as social media monitoring. In this paper we therefore explore the accuracy and applications of sentiment analysis and consider it against the related technologies on the market.

To this end a number of interviews and data comparisons were conducted. A number of sentiment analysis tools were chosen for the test, specifically The Stanford Classifier, Sentistrength, Semantria, Radian6 and “The Feeling Meter” from Copenhagen Business School. A test set of approximately 18500 Facebook posts were extracted via Radian6 and run against all of the tools, in order to be able to compare their performance. Furthermore, a representative sample of 1000 posts were extracted and manually annotated by four individual human coders. In further extension of the interviews, through discussion and speculations, a number of applications are suggested as opportunity for development or value generation. Due to the paper not focusing on a specific industry or application, it has not been possible to categorize value in specific terms.

The results showed that the commercial tools Sentistrength and Semantria were by far the most accurate, when compared to the human annotation. It was found that the tools are highly reliant on the data they have been trained on, as well as the training data plays a large role in domain adaptation. On the individual post level, all the tools performed extremely poorly, but all did see a significantly more accurate result (based on the human annotation) of considering the aggregate score. We conclude that the performance of the tools, might not be due to the technology, but rather the approach to sentiment that has been chosen in the industry. Labeling with 3-5 labels of sentiment only provides very minor value to the data, in order for it to have an valuable business application. Instead, we suggest an approach where sentiment is represented by a more dynamic numeric value, derived via annotators using a dynamic slider rather than a negative/positive label. This will create a different measurement for accuracy, on that might be more in tune with the actual annotation.

Through our interviews with a number of academics as well as Trustpilot and Falcon Social, we also speculate and discuss the extended value of using sentiment analysis as an extension of other technologies or for other applications. The conclusions, due to the low accuracy derived from our tests and through points raised in the interviews, is that the current form of sentiment does not utilize the potential of the data. We therefore suggest a
number of different extended applications, whilst the silver lining is the conclusion that sentiment analysis in its current form is uninformative, and needs to be augmented either with external data parameters, such as extralinguistic data, profile data or topic mining, in order to add more value. While extra dimension added in the Feeling Meter (arousal) adds a notable degree of extended information, the manner of which people express themselves on social media, means that insights beyond just sentiment, needs to be mined in order to gain appropriate and useful insights into the dialogue.
2.0 Table of contents

1.0 Executive Summary .................................................................................................................. 2
2.0 Table of contents ..................................................................................................................... 4
3.0 Introduction ............................................................................................................................. 8
   3.1 Research Question ................................................................................................................ 8
   3.2 Introduction .......................................................................................................................... 8
4.0 Methodological considerations ............................................................................................... 9
   4.1 Secondary Data .................................................................................................................... 9
      4.1.1 Literature reviews, framework .................................................................................... 9
   4.2 Primary Data ......................................................................................................................... 10
5.0 Technologies, terms and processes ...................................................................................... 10
   5.1 What is computational linguistics? ..................................................................................... 10
      5.1.1 Natural Language Processing ...................................................................................... 11
      5.1.2 Syntactic and Semantic analysis ................................................................................ 11
      5.1.3 N-grams ....................................................................................................................... 12
      5.1.4 Google n-gram viewer ............................................................................................... 12
      5.1.5 Ambiguity .................................................................................................................... 13
   5.2 Machine Learning ................................................................................................................. 14
      5.2.1 Supervised and unsupervised machine learning .......................................................... 14
      5.2.2 Multinomial Naive Bayes ........................................................................................... 15
   5.3 Stemming and lemmatization .............................................................................................. 16
   5.4 Stopword removal ................................................................................................................. 17
      5.4.1 Applications of stemming and lemmatization .............................................................. 18
   5.5 POS Tagging ......................................................................................................................... 19
      5.5.1 Stochastic POS tagging ............................................................................................. 19
      5.5.2 Unsupervised Tagging ............................................................................................... 19
      5.5.3 Output and disambiguation ......................................................................................... 20
      5.5.4 Application and insights ............................................................................................ 21
5.6 Sentiment analysis (Opinion mining) ........................................................................... 21
  5.6.1 Subjectivity and Objectivity .................................................................................. 22
  5.6.2 Degree of polarity ............................................................................................... 23
  5.6.3 Syntactic context: The Stanford Classifier ......................................................... 23
  5.6.4 Polarity, confidence level .................................................................................... 24
5.7 Topic identification, question answering, summarization. ......................................... 24
5.8 Shallow semantic parsing............................................................................................ 25
  5.8.1 Ambiguity: ......................................................................................................... 27
  5.8.2 Chunks and clauses ............................................................................................ 27
  5.8.3 Ambiguity in language ....................................................................................... 28
6.0 Current business applications, systems & technology .................................................. 29
  6.1 Translation softwares .............................................................................................. 30
  6.2 Digital personal assistants ..................................................................................... 31
  6.3 Watson .................................................................................................................... 31
7.0 Comparison of sentiment tools; findings and process .................................................... 32
  7.1 Tools....................................................................................................................... 32
    7.1.1 Radian6 ............................................................................................................ 32
    7.1.2 Sentistrength .................................................................................................... 33
    7.1.3 Stanford sentiment ......................................................................................... 34
    7.1.4 Feeling meter ................................................................................................... 35
    7.1.5 Semantria ........................................................................................................ 35
  7.2 Research Design....................................................................................................... 36
    7.2.1 Data source ....................................................................................................... 36
    7.2.2 Aim, credibility ................................................................................................ 37
    7.2.3 Method ............................................................................................................. 37
    7.2.4 Process description ......................................................................................... 37
    7.2.5 Pruning and normalizing text ........................................................................ 39
    7.2.6 Annotation....................................................................................................... 39
  7.3 Interview setup ........................................................................................................ 40
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.6.1</td>
<td>Syntactical and semantic complexities in annotation</td>
<td>71</td>
</tr>
<tr>
<td>9.6.2</td>
<td>Poor grammar, complex syntactic structure</td>
<td>71</td>
</tr>
<tr>
<td>9.6.3</td>
<td>Irony/Sarcasm, jokes</td>
<td>72</td>
</tr>
<tr>
<td>9.6.4</td>
<td>Opinion</td>
<td>72</td>
</tr>
<tr>
<td>9.6.5</td>
<td>Object of analysis</td>
<td>73</td>
</tr>
<tr>
<td>9.6.6</td>
<td>Context, continuity and emotional contagion</td>
<td>74</td>
</tr>
<tr>
<td>9.6.7</td>
<td>Shelf life of words, domain adaptation</td>
<td>75</td>
</tr>
<tr>
<td>9.6.8</td>
<td>Prior knowledge of systems</td>
<td>76</td>
</tr>
<tr>
<td>9.7</td>
<td>Interviews</td>
<td>77</td>
</tr>
<tr>
<td>9.7.1</td>
<td>Emotional arrays, valence and salience - the question of two-dimensional sentiment analysis</td>
<td>77</td>
</tr>
<tr>
<td>9.7.2</td>
<td>Mood, Emotions and feelings - multiple opinions</td>
<td>79</td>
</tr>
<tr>
<td>10</td>
<td>Discussion</td>
<td>81</td>
</tr>
<tr>
<td>10.1</td>
<td>Is the current level of sentiment analysis a viable, sustainable and informative approach from a business perspective?</td>
<td>81</td>
</tr>
<tr>
<td>10.1.1</td>
<td>Is sentiment sufficiently accurate?</td>
<td>81</td>
</tr>
<tr>
<td>10.1.2</td>
<td>Common pitfalls - where are the systems ineffective?</td>
<td>82</td>
</tr>
<tr>
<td>10.1.3</td>
<td>Falcon Social &amp; Trustpilot - Benefits and criticism</td>
<td>83</td>
</tr>
<tr>
<td>10.1.4</td>
<td>Is the current technical form informative?</td>
<td>84</td>
</tr>
<tr>
<td>10.2</td>
<td>What are the future areas for NLP and sentiment analysis?</td>
<td>89</td>
</tr>
<tr>
<td>10.2.1</td>
<td>Extralinguistic data - Profiles and contexts</td>
<td>89</td>
</tr>
<tr>
<td>10.2.2</td>
<td>Profiling - Personality Insights from text</td>
<td>90</td>
</tr>
<tr>
<td>10.2.3</td>
<td>Temporal, geospatial reasoning - determining intentions</td>
<td>91</td>
</tr>
<tr>
<td>10.2.4</td>
<td>Customer service</td>
<td>91</td>
</tr>
<tr>
<td>10.2.5</td>
<td>Comment moderation and the Asshole Meter</td>
<td>91</td>
</tr>
<tr>
<td>10.2.6</td>
<td>Intelligence</td>
<td>92</td>
</tr>
<tr>
<td>10.2.7</td>
<td>Cookie enrichment</td>
<td>92</td>
</tr>
<tr>
<td>10.2.8</td>
<td>Detecting complex dependencies in rich text</td>
<td>93</td>
</tr>
<tr>
<td>11</td>
<td>Ethics</td>
<td>94</td>
</tr>
</tbody>
</table>
Conclusion ................................................................................................................................................. 95

12.1 Sentiment analysis is inaccurate and uninformative ................................................................. 95
12.2 Sentiment is not informative on its own .................................................................................. 95
12.3 Future applications - which technologies are essential for a working system .......... 95
12.4 What industries/fields will it be seen in .................................................................................... 96

13 References .......................................................................................................................................... 97

Appendices start at page 102 and are not indexed

3.0 Introduction

3.1 Research Question

Does sentiment analysis present a value to business in both its current and future form?

3.2 Introduction

Sentiment analysis is a well-established Natural language processing (NLP) technology in business. Found in prominent social media monitoring tools such as Falcon Social’s listening platform\(^1\) and in studies such as Predicting iphone sales from iphone tweets (Lassen, Madsen, Vatrapu 2014). But how well does it work? And what insights does it actually provide, beyond a basic social media monitoring parameter? As we will explore in this paper, sentiment analysis is a fundamental theory, characterized by many approaches - many which struggle to provide an accurate representation of reality. We also seek to determine how sentiment analysis has a much greater potential, when coupled with wider NLP technologies and extralinguistic information.

We explore the accuracy of sentiment analysis through several market technologies such as Radian6, Semantria and Sentistrength and compare them to a small set of human annotators. We support our findings and expand our considerations of technologies and movements in the field through a series of interviews with a number of academics in the field as well as two business with Falcon Social and Trustpilot. Finally, through discussions and speculation, we outline a series of applications for sentiment and NLP technologies, based on our own knowledge and considerations of the field.

Considering viability and the current application of sentiment analysis, we seek to couple the link between business and technology.

\(^1\) https://www.falconsocial.com/
4.0 Methodological considerations

In order to apply a valid method to our study, data collection methods have been considered up against current development of the language technology field. Characterizing our topic, is the intention to evaluate sentiment analysis in its current form, uncover the potential business value of the existing technologies and *structuralize or categorize* the findings.

Business application, without any delimitation, can encompass both a general value and utilization, but can also extend into benefits that are industry specific. This broadness and heterogeneity of both the market (non-industry specific) and the technologies (ranging from sentiment analysis to shallow semantic labelling) prompts an exploratory approach with both a qualitative and quantitative data collection.

An initial round of exploratory interviews, supported by a literature review, provides insights into relevant topics and current assumptions of the performance of sentiment analysis systems and opening up areas of relevance. This is further complemented by a quantitative data collection and comparison through various systems. Said datasource and systems are described later in the paper, in research design (Section 7.2).

Evaluating whether the value generation is industry specific or can be generalized are due to the scope of the paper difficult to establish. It is necessary to perform at least one iteration to ensure the study supplies sufficient data, meanwhile this paper has focused on width rather than depth of its data analysis to adequately justify our hermeneutic approach (Fuglsang & Olsen 2004). The aim is therefore rather to establish premises for further, confirmatory studies of the assumptions derived in this paper.

**4.1 Secondary Data**

Secondary data is derived from previous courses and readings related to the field. Papers and leads are gathered throughout the process, and a literature review was conducted in the first stage of the process in order to identify relevant topics and outline the frameworks of the paper.

**4.1.1 Literature reviews, framework**

The literature review was conducted using an initial list of 30 different keywords, covering the broad and introductory terms such as *NLP, Computational Semantics, Computational*
Linguistics as well as well-known and specific technical terms such as parse trees, POS taggers, temporal reasoning. The queries led to an array of approximately 40 papers, which was screened for topics. A total of 14 topics and 28 subtopics were identified, and the papers were subsequently tagged with those relevant. This has in turn provided an overview for easy structure, and allowed a delimitation of topic, easing the identification of further papers.

The frameworks established drew an image of a field with many fragmented approaches. While there are a general consensus about a stochastic/probabilistic approach and several theories have achieved recognition, numerous papers are still occupied with improving methods. Though systems such as POS taggers have achieved a high accuracy, the high diversity within the generally composite technologies such as sentiment analysis systems suggests that there are still multiple unaddressed issues, in regards to achieving a reliable accuracy, that are not domain specific.

4.2 Primary Data

The method for collecting primary data has been structured on the ambition to determine the relevant technologies and movements within the academic field, identifying current applications and testing the accuracy and concordance of various systems. Due to the twofold orientation - current viability of sentiment analysis and the future applications within the field - both a quantitative data collection is conducted as well as a qualitative through interviews.

This is described further in section 7.2 - Research Design.

5.0 Technologies, terms and processes

5.1 What is computational linguistics?

The field of computational linguistics is occupied with natural language processing (NLP), utilizing a computational approach to retrieve data and deduct information from text. Said data may be syntactic or semantic representation and interpretation of grammar and language. Further expanding the concept, it has ties to psycholinguistics and sociolinguistics in regards to the the application of said data.
Due to the complex nature of the field, it draws on experts from numerous fields such as computer scientists, logicians, mathematicians, etc. The field is closely tied to artificial intelligence, and is in many aspects founded in the 1950’es automatic translation efforts. Early work was based on a rule-based representation of grammar, but has in the recent years shifted towards a Stochastic (probabilistic) approach and hybrid forms.

5.1.1 Natural Language Processing

NLP within computational linguistics works well within the same sphere of AI and computer science. The field however focuses on the human-computer interaction of languages, typically on how computers can derive meaning from natural human language.

Similar to computational linguistics, most research has been founded in the stochastic school since the 1980 using technologies such as machine learning as computational power has increased rapidly and the influence of Chomskyan theories of linguistics receded.

Much understanding of natural language is reliant on parsing human language into a structure, which can then later be attributed certain values. POS taggers, for example, allows for a syntactic construction of sentences, allowing researchers to identify topic or couple it with other technologies to identify sentiment or similar analysis.

5.1.2 Syntactic and Semantic analysis

Within the numerous technologies associated with NLP, some are merely pruning techniques, whereas others offer a degree of analysis. Syntactic analysis is the most fundamental aspect in terms of analysis, offering a contextual representation of the language. This type of analysis forms the foundation for many question answering and topic identification systems. As it also considers the relation between words, it also tends to have a role in e.g. sentiment analysis, where it will be capable of identifying words working in e.g. intensifiers and negators of other words based on their relation.

Semantic analysis is somewhat more complicated, and the line between syntactic and semantic analysis is not necessarily entirely clear. However, semantics often work with understanding the inherent meaning deriving e.g. intention of a statement. Shallow semantic parsing, for example, illustrates some aspects of this in terms of identifying the agent, object
and action in order to derive more meaning from the text. However, this semantic analysis is significantly more complicated and highly reliant on various lexicalized and annotated word and treebanks, as well as not nearly as accurate as some of the established methods of syntactic analysis.

5.1.3 N-grams
The statistical approach to NLP is one of the most widely used ways of learning and predicting language. One of the main techniques used for the statistical approach is known as N-grams. N-grams are mathematical representations of how often a combination of words (the n-gram) appears in a corpus. This method can be used to predict or guess the most likely word to follow a phrase, which is especially practical when interpreting input from speech (noisy channel) (Brown, 1992).

5.1.4 Google n-gram viewer
N-grams are used for language modelling in several cases of NLP. The applications varies from machine translation, over spell correction to speech recognition.

As part of their ‘Books’ project, Google have created a tool called the n-gram viewer, that illustrates the basics of n-grams.

![Comparison of child care (green), kindergarten (red) and nursery school (blue)](https://books.google.com/ingrams/info)

Fig. 01 - Comparison of child care (green), kindergarten (red) and nursery school (blue)
By specifying search terms, the user can see the over-time development of occurrences of a certain word or phrase, and even compare these to other words or phrases. As seen in figure 01 (above), the word ‘kindergarten’ has been fairly steadily used since 1950, whereas the bi-gram child care has become gradually more popular since the late 60’s. The n-gram viewer uses the Google Books’ corpora of literature, which with English literature goes back to the 1500’s.

As previously mentioned, n-grams can be used to help determine what was said when analysing language from a ‘noisy channel’ – e.g. speech. When doing speech-to-text analysis it can be difficult to know whether the person said, “world’s end” or “world send”. As a thinking human being we can easily make a qualified guess that the person probably meant “world's end”, but how do we define this as a rule in an algorithm? This is where n-grams become useful as we can make guesses based on statistics:

![Figure 2](https://books.google.com/ngrams/)

Looking at the corpora for the years 1990-2000, we see that the phrase ‘world's end’ has occurred significantly more frequently than ‘world send’, and our algorithm now has a basis for guessing which of the two interpretations were correct.

### 5.1.5 Ambiguity

Another way of using n-grams can be to handle ambiguous words when analyzing text. Ambiguosity is one of the biggest challenges in NLP, as it can be very difficult to automatically detect which meaning of a word is the correct in a specific context. In this
case, n-grams can be used to give a qualified estimate of which meaning is relevant in the specified context based on how often that meaning has appeared in similar contexts. (Sidorov et. al., 2014)

### 5.2 Machine Learning

Machine learning within NLP or Computational Linguistics is occupied with recognizing and identifying patterns within text. The aim is to allow the system to learn, while not being explicitly programmed. (Arthur Samuel, 1959). As phrased by Tom Mitchell from Carnegie Mellon University:

> "A computer program is said to learn from experience $E$ with respect to some task $T$ and some performance measure $P$, if its performance on $T$, as measured by $P$, improves with experience $E$."

*(Mitchell 1997)*

One such example within NLP is spam classifiers. Reliant on calculating the probability of a certain words indicating that a message is considered spam, it learns based on recognizing which patterns are found in spam and non-spam, thus improving the effectiveness of the filter.

#### 5.2.1 Supervised and unsupervised machine learning.

Within the field of machine learning, there is a clear distinction between unsupervised and supervised machine learning. Supervised machine learning is trained on a pre-defined or pre-annotated training data set. The conclusions or results drawn from this data is intended to assists the system in reaching the correct or most accurate conclusion when given new data.
Unsupervised machine learning, on the other hand, is simply given a large amount of data, and must itself find patterns and relationships therein. One example is clustering where the system clusters words based on how often they occur in relation to each other. Contrary to supervised ML that often try to yield a predictive results, unsupervised ML is often used for descriptive modelling, identifying common or uncommon datapoints (Jain, 2015).

A simple predictor for supervised machine learning, may look like this (McCrea, n.d.):

$$h(x) = \theta_0 + \theta_1 x$$

Where $\theta_0$ and $\theta_1$ are constants, and $h(x)$ is the predictor function - $x$ being the given input data and $h(x)$ being the prediction on when given the specific value. Optimizing $h(x)$ is typically done using training examples, yielding a result known in advance. Comparing the correct value and the predicted value enables us to compare the output between the two, a tweaking the constants, until the system is optimized to be accurate enough, to be applied to real-world data. However, it should be noted that most ML systems are far more complex than this, utilizing several more dimensions and polynomial terms. Some systems even rely on thousands or millions of dimensions of data to yield an accurate prediction.

This also illustrates a linear function, whereas many systems will be based on something non-linear. Many machine learning systems are based on a Naive Bayes Multinomial classifier, or a derivative thereof often yielding a probability between 0 and 1 of how likely a given instance is.

### 5.2.2 Multinomial Naive Bayes

Multinomial Naive Bayes classification, also known as just Naive Bayes, is the mathematical model for calculating the probability of a word being in one category or another. It builds on a statistical approach where the probability of a word being in one category over the other is calculated based on statistics from a training set. The math is as follows, and will be analyzed in-depth further on.

$$\phi_k|y=1 = p(x_j = k|y = 1) = \frac{(\sum_{i=1}^{m} \sum_{j=1}^{n_i} 1\{x_j^{(i)} = k \text{ and } y^{(i)} = 1\}) + 1}{(\sum_{i=1}^{m} 1\{y^{(i)} = 1\} n_i) + |V|}$$
Although seemingly complicated, the math behind Naive Bayes is not that difficult to understand. The example above, tries to determine the probability of an e-mail being spam or not.

The data used for the classification is based on a frequency count of a larger set of manually classified e-mails. Through frequency counting it is determined how many occurrences is of every single word in the entire dataset, which is then called the dictionary. To determine the probability of an e-mail being spam, we look at each word of the email in turn.

The first formula (above) calculates the probability of a single word being a “spam-word” or not. The answer comes from seeing how many times the word occurs in a spam email, divided by how many times the word occurs in a the training set in total plus the total number of unique words (the vocabulary). The verdict then comes from multiplying the probabilities of all the words in the particular e-mail, being either spam or not-spam, and then we end up with a number, where the category with the higher number is the most likely for our text, therefore it is highly reliant on the size of the training set.

\[
\phi_{k|y=0} = p(x_j = k | y = 0) = \frac{\left( \sum_{i=1}^{m} \sum_{j=1}^{n_i} 1\{x_j^{(i)} = k \text{ and } y^{(i)} = 0\} \right) + 1}{(\sum_{i=1}^{m} 1\{y^{(i)} = 1\} n_i) + |V|}
\]

\[
\phi_{y} = p(y = 1) = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}{m}
\]

(McCrea, n.d.)

5.3 Stemming and lemmatization.

Stemming and lemmatization is an integral part of many NLP technologies as it seeks to reduce inflectional forms and derivationally related forms of a word into the base form of the word (Manning et al 2008). While stemming refers to a crude, heuristic approach removing the endings, lemmatization refers to returning a words to it’s lemma by using a vocabulary and morphological analysis.

Stemmers, such as the Porter Algorithm, are based on rules, such as converting endings “sses” to “ss” as with caresses to caress. Later rules also take in consideration the syllables,
to consider whether the suffix is part of the stem of the word, and thus shouldn’t be removed.
As Manning et al. (2008) explains:

“For example, the rule: (m>1) ement -> would map replacement to replac, but not cement to c”

Manning (2008)

Hence, stemmers are smaller and simpler from a programming point of view and has a
higher recall than lemmatizers, but lacks in precision, as it would stem operative and
operating to oper, which could complicate further information extraction, especially in n-gram
contexts, such as operating theater.

Lemmatizers may have less recall, but allows for greater precision, e.g. better being
lemmatized to good, which would have been missed by stemmers as this requires a
dictionary lookup. Similarly, lemmatizers may use a treebank where the words have been
assigned semantic meaning as in separating a river bank from a money bank.

However, choosing between lemmatizing and stemming is a question of the further
application. In many aspects, stemmers will perform sufficiently. Accuracy of lemmatization
also depends on the dictionary it is tied to and possibly whether the text is POS tagged in
order to discern between nouns and verb forms of the same word, as it is with CST’s
lemmatizer2. While it may be more accurate in some instances:

He operated on the operative within the operating theater:

Porters stemmer3: He oper on the oper within the oper theater

CST’s lemmatizer: He operate on the operative within the operate theatre

5.4 Stopword removal

Stopword removal is, similar to stemming and lemmatization, a preprocessing task in many
NLP systems. It is used in a variety of different tasks such as Unsupervised clustering:
Removing stopwords before creating clusters as the high presence of stopwords would
influence the clusters by their presence.

2 via http://cst.dk/online/lemmatiser/uk/
3 via http://text-processing.com/demo/stem/
In **supervised machine learning** for removing stop words from the feature space.

In **information retrieval**, ensuring that stop words aren’t indexed

In **text summarization softwares**, to ensure that stopwords do not count towards text summarization scores or are identified as topics. (Ganesan n.d.)

Depending on the application of the stopword removal, various types of words can be classified as stopwords. Most commonly *a, an, the* (determiners) will be removed, but others such as *above, adjacent* (prepositions) or *good, bad* (adjectives) are also considered stop words in some cases. However, in other cases these can potentially be detrimental to the process. For example, in geospatial reasoning prepositions may give clues to location, whereas adjectives will play a considerable part in sentiment analysis. Similarly, there may be domain specific stop word removal systems, e.g. in the medical industry, where certain abbreviations are not considered relevant to the context. This also goes for social text, which may be characterized by a significantly different language use, than academic texts.

Similarly, some terms or searches are disproportionately affect. *A flight to New York* may lose a large part of its meaning, by removing the preposition *to*. The tendency of information retrieval systems have been a decreasing size of stopword banks, from 200-300 down to 7-12 words, where some systems do not utilize stopword removal at all.

As it is clear, it does recognize a distinction between operate and operative. However, it still does not consider n-grams as entities such as operating theater.

### 5.4.1 Applications of stemming and lemmatization

Applications vary but they are an integral part of especially spam filters in order to stem or lemmatize words, in order to increase detection for a Naive Bayes classifier. It has also been used in search engines with differing success. It was implemented by Google in 2003, that words such as *working* would yield results for *work*, however the algorithms have be refined significantly since. Lemmatization is seen in similar contexts, as well as in topic identification and information retrieval.
5.5 POS Tagging

Part of Speech (POS) taggers are tools for marking up or tagging words based on their definition and context, i.e. relationships with surrounding words within the same sentence or paragraph. It is often used as preprocessing for parsing as it reduces the number of parses (Nau, 2010). The most basic tags relate to word types; Noun, Verb, Adjective, Adverb, preposition (in english grammar) while other systems, such as the UPenn TreeBankII has 36 different tags4.

5.5.1 Stochastic POS tagging

POS tagging typically follow two different approaches. While there are many different systems, ranging from lexicalized datasets to hidden Markov models, they can mainly be divided into a rule-based and stochastic (probabilistic) approach. Most newer research is reliant on stochastic models, hence this will be the focus (Manning & Schütze 1999).

The stochastic approach relies on a pre-tagged training set, such as the Brown corpus5 or the aforementioned UPenn Treebank. It subsequently relies on Bayesian probability calculation, on whether a word in the given sequence, has probability (based on the training set) to be T tag (N, V, adv., etc).

\[
P(T \mid W) = \frac{P(W \mid T) P(T)}{P(W)} = \alpha P(W \mid T) P(T)
\]

With maximum entropy classifiers as an alternative to Naïve Bayes, it is possible to achieve an accuracy of more than 95% using the UPenn Treebank (Toutanova & Manning 2000), underpinning that many POS taggers are considerably accurate. Maximum entropy classifiers may require more time to train and optimize but does in turn provide better results as well as good performance in regards to CPU and memory consumption (Vryniotis, 2013)

5.5.2 Unsupervised Tagging

It is therefore also possible to use unsupervised taggers also heavily reliant on machine learning utilizing an unlexicalized corpus. As the texts are not annotated, it relies purely on statistics to detect certain patterns in the use of words; e.g. an occurring rather than a in

---

4 https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
5 http://clu.uni.no/icame/manuals/
certain contexts. While it may not be able to establish a formulated rule, it may be able to establish patterns which can also give some insight in the language, and may be a solution to domain adaptation difficulties as it is not dependent on a certain linguistic structure or syntax.

While current development favors probabilistic methods, such as the unsupervised tagging over rule based, there are also examples of hybrid or ensemble methods, e.g. Brill’s transformation based tagger (Brill, 1995). While this was formative for many systems, it seems to have been eclipsed by state of the art systems such as maximum entropy classifiers, achieving higher accuracy.

5.5.3 Output and disambiguation

Using University of Illinois POS tagger⁶, will yield following result, based on the UPenn Treebank⁶:

```
DT The JJ quick JJ brown NN fox VBD jumped IN over DT the JJ lazy NN dog
```

- DT - Determiner // JJ - Adjective // NN - Noun, singular or mass // VBN - Verb, past participle // IN - Preposition or subordinating conjunction

As it is visible, the tagger is capable of determining a large array of word classes and tense. As mentioned earlier some taggers achieve accuracy up around 95%. Similar, the UIP POS tagger also illustrates the capability of dealing with disambiguation which was an issue in the earlier systems.

```
PRP He MD would VB object IN towards DT the NN use IN of DT a JJ certain NN object
```

As illustrated the system recognizes the (syntactic) difference between object and object. Of course, this is not representative of the entire system, but it serves as an example of the capabilities and reliability of POS taggers.

⁶ http://cogcomp.cs.illinois.edu/page/demo_view/POS
While it is capable of recognizing differences in the use of words, POS taggers are still challenged in regards to semantic complications of homonomy, synonymy, polysemy, hyponymy and hypernymy (Section 5.8.3; Abbot 2013)

The above example, shows how the POS tagger does not yield an understanding of *big sister* as an entity. While this is, partially, addressed through n-gram models and wordnet apps, it is later discussed how semantic role labelling performs better in a semantic context.

**5.5.4 Application and insights**

POS tagging is typically an important aspect of NLP technologies, especially sentiment analysis, as it is fundamental for the breakdown of the syntactic structure of the text. The insights provided are therefore also related to its purely syntactic capabilities. I.e., it may let us know that the noun *coverage* is preceded by the adjective *bad*.

POS taggers are essential for word-sense disambiguation and are also typically seen in question answering systems and information extraction.

**5.6 Sentiment analysis (Opinion mining)**

The technology of sentiment analysis refers to the use of NLP, Computational Linguistics and text analysis to identify the sentiment expressed in a given text, often scored within values for negative, neutral and positive attitude, i.e. determining *attitude* and *polarity* of text. Applications are often seen in relation to social studies or marketing (Pak & Paroubek 2010).

There are significant variations in how sentiment analysis systems are built as they can be considered composite systems (combining e.g. POS taggers, lemmatizers etc). Most rely on keyword spotting and lexical affinity as well as looking at grammatical dependencies, through deep parsing of the text; for example when considering intensification and negation. Many tools will utilize previously mentioned tools and technologies, such as tokenization, lemmatization, stemming, stopword removal to “prime” the text prior to applying an n-gram based Naive Bayes classifier to the text, similar to what it seen with spam classifiers. Central

---

to the techniques are also the application of Machine learning techniques (Section 5.2), where a system is trained to evaluate and score text segments.

However, the composition of the approach may vary. Some systems are reliant on supervised learning, utilizing large corpora of lexicalized polarized words, e.g. Esuli & Sebastiani (2005,2006), while others construct their systems for a semi-supervised or weakly supervised setting, using only a handful of seed words (Hassan et al. 2014). Similarly, the latter utilizes a Markov Random-Walk to establish relation between words based on few words with an indexed polarity, thus providing an example of weakly supervised learning, while still relying on Wordnet for annotations and word-relations.

5.6.1 Subjectivity and Objectivity
Pang & Lee (2004) utilize an approach reliant on identifying subjective and objective sentences, discarding the latter, to achieve a greater accuracy in determining opinion as “irrelevant sentences” are no longer considered.

“...machine-learning classifier to the resulting extract. This can prevent the polarity classifier from considering irrelevant or even potentially misleading text: for example, although the sentence “The protagonist tries to protect her good name” contains the word “good”, it tells us nothing about the author’s opinion and in fact could well be embedded in a negative movie review. “

Pang & Lee 2004

A similar distinguishing is seen in Pak & Paroubek's (2010) work; separating the objective and subjective text (sentences) before attempting to extract opinions from the dataset, however their work also exemplifies adapting the technique to the appropriate dataset, using emoticons inherent in the twitter corpus they are examining as an indicator for general sentiment. Pak and Paroubek (2010) note that POS tags can also be an indicator between positive and neutral sentiment as they are used differently in their datasets with a strong presence of proper nouns in objective texts, while subjective texts tend to be characterized by a stronger presence of personal pronouns. Similarly, they observe use of third person in objective texts, and first and second person (the audience) in subjective texts. Therefore, stochastic forms of sentiment analysis, can benefit from some rules, e.g. in relation to POS taggers, in order to achieve a higher accuracy.
5.6.2 Degree of polarity

There are various emerging approaches to conducting sentiment analysis. A traditional approach is via a lexicalized dataset for positive and negative words. Scoring only on a 1-gram basis, as some earlier systems are known to, does not yield reliable data. Instead a series of different indicators are needed for a reliable result such as the syntactic context taking into consideration negation, intensification and contrast. Negation, as the word suggests, refers to a another word negating the following. E.g. “The system did not perform well” whereas intensification refers to words, strengthening the statement “The system performed very well.” and contrast: “The system had an awful UI, but I still liked it”.

Of course, these may all be present, thus demanding more of the system “The system was not very good, but I liked the UI".

Not only does this complicate the issue of negation, intensification and contrast, it also raises some issues in regard to what are identified as the statements about the topic (the system) and what is the actual opinion of the author, which is briefly touched upon by Pang & Lee (2004) but not explored as separate entities.

5.6.3 Syntactic context: The Stanford Classifier

Negation and intensification are easily visualized using the Stanford classifier. Resembling the tree structure of a POS-tagger or a Parse Tree, it deconstructs the sentence(s) and orders it hierarchically based on the sentence structure. Adverbs such as very are therefore treated as an intensifier, increasing the assigned value of a following statement.

![Figure 1: Example of the Recursive Neural Tensor Network accurately predicting 5 sentiment classes, very negative to very positive (→, ←, 0, +, ++), at every node of a parse tree and capturing the negation and its scope in this sentence.](image)

*Figure 3, retrieved from Socher et al. 2013*
5.6.4 Polarity, confidence level
Certain sentiment systems will also calculate a confidence level indicating how certain it is that a statement is, in fact, negative/positive. This confidence level may or may not be considered the same as the polarity of the statement - how strongly positive or negative the statement is. Other systems, such as the Stanford Classifier uses negative and very negative as labels, due to the intensification and negation techniques incorporated in their recursive neural model. However, determining the intensification or polarity of the statement could be perceived as a useful piece of information in regard to ambiguous or contradictory.

Some posts may include a quote and the author then relating a statement to that quote. Such a post would be very polar and provide a somewhat small confidence level, or in some systems, yield a neutral post as the two different statements would cancel each other out.

5.7 Topic identification, question answering, summarization.

Topic identification in NLP is the process of automatically identifying and extracting topics of larger amounts of text. A very coarse simplification of the technique used to do the identification, is that the system counts the frequency of all words in the corpus and classifies the most frequent words as main topics.

Topic identification can be used for several different purposes. E.g. a customer service department might want to have all of their e-mails analyzed to see which topics people usually ask questions about.

Related to this, Chali & Hasan (2015) have done work on using topic identification to improve on question answering. One of the issues with automated question answering is understanding complex questions, especially when asked about more than one thing. By using topic identification on a complex question, it is possible to break down the question into relevant topics and automatically generate simpler questions to be answered, hoping that the combination of different simple answers can answer the more complex question.

We have previously mentioned that topic identification is done by counting the frequency of the words in the corpora and letting the most frequent words be a representation of the topics. Of course, for this method to make sense, we first have to remove the stopwords (section 5.4) in order for words like “and” to not be categorized as a topic. At the same time a stemming or lemmatization process has to be carried out on the corpora to make sure that words like “running” and “run” will be categorized together as the same topic.
5.8 Shallow semantic parsing

While a POS tagger enables syntactic parsing or deconstruction of text, shallow semantic parsing and semantic role labelling attempts to apply a semantic understanding/interpretation of text. Semantic role labelling identifies word-roles such as agent, patient, source and destination. (University of Texas, n.d.)

A number of published papers on shallow semantic parsing describe utilizing FrameNet (Gildea & Jurafsky 2002, Coppola et al 2009 to mention some) which is a lexical database containing more than 170,000 manually annotated sentences, both machine- and human readable. It works in so called frames (e.g. cooking) utilizing frame elements (cook, food, container) and words that invoke this frame; lexical units (fry, bake, boil) (About Framenet, n.d.). Using machine learning to extract syntactic features and annotated roles, it is possible to construct a large set of rules for syntactic analysis for semantic frames. However, while this database provides an understanding of a lexical structure of the frame it does not contain the rich information on a word level basis that WordNet provides (Shi, 2004).

“Syntactic relations can be described using a set of rules about how a sentence string is formally generated using word strings. Instead, semantic relations between semantic constituents depend on our understanding of the world which is across languages and syntax”

(Shi, 2004)

The inherent data for NLP purposes is tremendous as both frameworks provide a significant collection of data when put together. As illustrated in Shi’s (2004) paper, the deconstruction using both frameworks allow for a datamining containing a significant information, addressing hyponymy and hypernymy as well outlining some speculations on deep linguistic analysis.

Other linguistic resources and knowledge bases are also combined in a later paper (Shi, 2005) which also provides some interesting aspects for deeper semantic analysis of text. While it is mentioned that it requires extensive human effort, there is still some success in using multiple databases (WordNet, VerbNet, FrameNet) to augment each other.
University of Illinois' online demo of their SRL tool, allows for illustration of the technique:

![Diagram](http://cogcomp.cs.illinois.edu/page/demo_view/14)

*Figure 4 - Snippet from http://cogcomp.cs.illinois.edu/page/demo_view/14*

Via this screen grab it is possible to identify several different processes. Lexicalization or a similar morphological analysis of the word is used to deduct jump from jumped. The *thing jumping* label is determined by the FrameNet and the syntactic analysis of the sentence is capable of clustering *the quick brown fox* as an entity and determine that it is the entity which is exerting the verb jumping. Similarly, the semantic analysis is capable of identifying that the object which is jumped over is the dog. However, it also illustrates a slight lack of accuracy, as the lazy dog is not recognized as an entity. Another modified example, illustrates the capability, but also the limitations of the semantic understanding to be tied to the annotation of in FrameNet:
Figure 5 - Snippet from http://cogcomp.cs.illinois.edu/page/demo_view/14

Capable of detecting the endpoint or the object of “interaction”, it shows a small degree of temporal and geospatial reasoning.

It is for this reason that Shallow Semantic Labelling or Shallow Role labelling are often used in question answering systems.

5.8.1 Ambiguation:
The word “borders” in a sentence might be tagged as a verb but also refer to a geo-location “Spain borders France and Portugal”. POS taggers are often able to identify whether or not the given word is used as a verb or a noun (etc.), however some work in the field of shallow semantic labelling works with the attempt to understand this a geo-location “Spain’s border with France”.

Similarly, it works on an multiple-words level categorizing topics as more than just the particular noun. As visible from the example, the semantic labelling tool from University of Illinois is capable of semantic role labelling as well as detecting negation (Roth, 2008, 2009, 2013).

5.8.2 Chunks and clauses
Shallow semantic labellers are typically reliant on POS taggers and identify syntactic structures called chunks and clauses. Clauses are the highest level where the text can be broken down into statements. Chunks identify the noun phrases, verb phrases, prepositions (etc) and thus further refine the parse tree to be more fine-grained. The predicate-argument structure builder, then uses this in addition to the previously mentioned wordbanks to
construct the predicate-argument structure. It must transcend entire statements as a reference such as “who”, in this case John Smith, must be referred back correctly.

Figure 6 - An example of a parse tree and verb-argument structure - Snippet from Punyakanok, Roth, Yih 2008

Semantic role labelling carries a considerable potential in regard to identifying topics, intentions, places and timeframes are tremendous however reliant on good grammar for the POS tagger and spelling for the (presumably) lemmatizer.

5.8.3 Ambiguity in language

Text ambiguity is one of the major challenges for many text parsing systems and can be challenging for both the syntactic and semantic parsing. Ambiguity typically refers to homonymy, synonymy, polysemy and hyponymy and provides a series of challenges (Abbot, 2013)
- Homonomy refers to the multiple meanings of a word: “the bank of the river” and “the bank next to Sears” both use the word bank, but hold different meanings. One of the challenges for a system, can be to distinguish this different meaning.

- Synonymy refers to similar words, synonyms, which can complicate specific situations, due to the slightly altered meaning, such as the difference between “Dan’s big sister” and “Dan’s large sister”.

- Polysemy refers to a different contextual meaning of the same word; e.g. “The human species (i.e., man vs. animal)”, “Males of the human species (i.e., man vs. woman)”, “Adult males of the human species (i.e., man vs. boy)”.

- Hyponymy and hypernymy refers to the hierarchical relationship between words; how gold and amber is subordinate to the word yellow.

While differentiation between these different meanings and words are addressed in POS tagging, it is still problematic to identify the distinctions beforehand. The sentence “I walked along the bank of the river, to get to my local bank branch.” will yield 2 occurrences of the word bank, even though bank has different meaning. This is typically mitigated by using Word Sense Disambiguation systems (WSDs) which relies on large corpora of unlabeled or annotated texts or knowledge based approaches with machine readable dictionaries and semantic networks (Kumar et al, 2012). While an unsupervised learning approach using machine learning is also possible, it works more as a clustering method which might complicate deriving particular meaning from the words / expression when further processing the outcome. (Hilborg & Nygaard 2014)

6.0 Current business applications, systems & technology

NLP technologies are found in a variety of technologies and already have numerous business application. Some applications are rather rudimentary, such as spam classifiers, while others are more complex such as sentiment analysis or translation softwares. The complexity varies but due to rapid development of computational power, many of systems are seeing availability to the wider public.

A number of well known technologies are spam classifiers, chatbots, text summarizers, translation softwares, sentiment analysis tools and to a wide extent, digital personal

---

8 Example retrieved from wikipedia: http://en.wikipedia.org/wiki/Polysemy
assistants such as Apple’s Siri. The emergence of such consumer products indicates that NLP has tremendous potential. Understanding the potential business application of NLP technologies such as sentiment analysis requires a consideration of the broader field of technologies as well as what is available on the market as of now.

Characterizing the most rudimentary systems such as chatbots and spam classifiers are the reliance on data. While some chatbots are rule based the most well known Cleverbot is simply based on accumulated input from thousands of users, ensuring that there are an appropriate reply. Spam classifiers are reliant on a large annotated dataset to ensure that it can calculate the the probability of whether a word is spam or not.

Text summarizers rely on counting frequency of words and calculating what is known as the rouge score based on n-grams to select the sentences addressing the topic, which will in turn be gathered and used for the summary. The systems are usually fairly simple and does not change the wording but simply removes a number of sentences.

6.1 Translation softwares

Among some of the most popular and widely known NLP tools are public translation tools. Machine translation has long been researched and tried perfected, but even the most advanced tools still leave a lot of room for improvement (Stix, 2006).

One of the most notable attempts at machine translation has been the Google Translate project. Google translate offers free online translation of text, speech and can even change the text in pictures to translate (Translate images, n.d.).

Google uses a stochastic model to translate texts based on a 100+ million document corpora (Translate, n.d.). The documents - e.g. officially translated documents from the U.N., are used to train a stochastic n-gram based translation model. By using n-grams it is possible to pick the word most likely to be correct in a sentence where there are several translations of the same word.

As the quality of machine translation tools is still questionable, it appears doubtful that it has been used in large scale applications. With the case of Google Translate, the tool has been built in to browsers (Google Chrome) to translate websites directly. However, for website owners (or others) who want to translate text automatically, Google offers a paid API service.
6.2 Digital personal assistants

As smartphones become “smarter” we have started seeing a growth in the distribution and usage of smart personal assistants, such as Siri (iOS), Google Now (Android) or Cortana (Windows Phone). Just lately Facebook announced their own personal assistant tool included in the Facebook messenger called M. M is a combined service using both NLP and some human interference.

Expect from M from Facebook these services focus on helping with simple tasks on your phone, like creating events or reminders, calling or texting someone or answering simple questions. The question answering is especially interesting in regard to topic identification, as previously discussed.

A service like Facebook M is interesting, because it combines the NLP capabilities of some of the already known services with a sort of concierge service. This allows Facebook to offer something that seems like a full-time personal assistant at a very low price as the services that require human interaction might be easier to bill the customer for.

6.3 Watson

The IBM Watson system is probably one of the most sophisticated implementations of NLP, and A.I in general, the world has seen so far. In short, it is an advanced question answering system that calculates and evaluates the answer most likely to be correct to any given question. From “reading” and analyzing humanly curated and selected data, Watson learns the language and domain specific elements of a new area and becomes “an expert” in the field (What is IBM Watson?, n.d.). Watson is then given a set of manually coded question pairs (questions and answers) within that field to further learn how to understand and answer certain question types. It then keeps evolving it’s knowledge by interacting with real people, getting some of its interactions rated by human experts (What is IBM Watson?, n.d.). Even though this is one of the most sophisticated NLP systems currently, we still see how much Watson relies on humans to curate, and help i adept.
7.0 Comparison of sentiment tools; findings and process

7.1 Tools

7.1.1 Radian6
Radian6 is a widely used commercial tool for collecting data and analyzing sentiment. However, it seems widely regarded as inaccurate and the provider have not disclosed any information on the structure of the system or how it performs its sentiment analysis. When researching Semantria we found that Radian6 is based on the same “Lexalytics Salience Engine” as developed by the company that runs Semantria (API Basics, 2015).

In order to determine accuracy and infer certain assumptions in the structure of the program, we conduct a number of tests to determine what, how and how well the system performs in regard to sentiment.

7.1.1.1 Prior Assumptions
Based on previous experience with Radian6, we examine the assumption that Radian6 works using a small lexicalized word-bank. The high percentage of neutral posts in the dataset suggests that it is unable to label these rather than them being ambiguous or objectively stated and hence that the sentiment is based on a small number of words and the absence of these will therefore yield a neutral result.

Furthermore, the tempo with which it allocates sentiment to the number of posts is unreasonably quick, but Falcon Social strives for a standard of 1000/sec (Alfredsen 2015) so it is unlikely that this is informative.

7.1.1.2 Setting up a data collection & sentiment analysis in Radian6
Radian6 is a fairly easy tool to use as its main customer group is non-technical sales and marketing people. It let’s the user specify “topic profiles” which define the search criteria for data that the user wants to analyze. This covers the search terms, media sources and sentiment topic. The topic profile is then used to create various dashboards that can help monitor the data aggregated from the different sources.

For the purpose of our tests, we set up a topic profile with various search terms for Donald Trump and with sentiment towards him. From hereon we were able to extract the data presented in the dashboards 10000 posts at a time.
7.1.3 How the tests are carried out

Seven datasets are collected in order to test the assumptions and the accuracy of Radian6. All datasets have the primary keyword *Trump* and one of the secondary keywords: President, Candidate, Republican, Donald and 2016 for the period September 1st - 15th. The object for the sentiment is then calculated by Radian6 towards each one of these secondary keywords, and the two additional words *Hillary* and *Cake*.

The sentiment calculation on *cake* and *Hillary* was chosen, in order to investigate if the system actually used the keyword as part of the analysis.

These posts are then compared when manually annotated to evaluate whether these relate to the topic or the overall sentiment of the sentence.

Secondly the reason for using the same dataset, but altering what topic the sentiment is directed at, is to determine if there are any obvious common denominators in terms of words - e.g. if the words fantastic, great, terrible etc. have a high frequency - and to evaluate the fluctuations in sentiment to further conclude on whether it is the post or the topic that is the determining factor. If it is easy to observe certain words has a common presence, it should theoretically be possible to reverse engineer the word-bank. This is not the aim of the study, but if a significant amount of the neutral posts contains a clear sentiment according to human annotation this is theoretically possible (at least to derive some insights).

7.1.2 Sentistrength

Sentistrength uses each individual sentence as object for analysis (similar to Stanford), then scores them and aggregates this on post level allowing for analysis of multi-sentence posts. It also scores the posts with an aggregate value ranging from -5 to +5, categorizing 0 and above as positive.

Accuracy of Sentistrength is estimated to 60.6% for positive feelings and 72.8% for negative feelings (Buckley et al 2010). It does however evaluate each sentence separately, then compares the sum of each sentence, yielding the average sentiment. Post with contradicting statements (posts including quotes) might therefore be a problem. The system is trained on social media data and applied to myspace during testing, where they argue that simple rules of spelling and grammar are not abided by (Grinter & Eldridge, 2003; Thurlow, 2003).
The system is based primarily on a word-bank of 298 positive terms and 465 negative terms, modified by a training algorithm to modify word strength. Similarly, the system also contains a spelling algorithm, booster word list, negative word list (word that invert emotion), repeated letters, emoticon list, repeated punctuation (intensifier at exclamation marks) and set to ignore negative emotion in questions (Buckley et al 2010).

The system overall has an interesting setup and is also not reliant on a POS tagger due to the system being aimed at forums which do not abide but such grammatical rules. Whether this has an improved accuracy in comparison to other systems is yet to be seen.

7.1.3 Stanford sentiment

The Stanford Classifier has been applied in a slightly different manner than the other tools. While the web-demo allows for multiple sentence level posts, the API access used in this paper to crunch large amounts of data does not. The system splits on breaks (dots, colons, question marks and expression marks), thus prompting us to test the influence on the various systems and see if the aggregate result is equivalent. The Stanford Classifier therefore yields a significantly higher number of units of analysis as each post is divided into sentences which are labelled as individual posts.

Issues with this structure, is that it may not consider the entirety of a post to be negated by a prior or following comment. However, assuming that it values each sentence individually (similar to Sentistrength described in the previous section) the aggregate distribution might be similar.

The Stanford classifier uses what has been dubbed a recursive neural model in order to deal effectively with intensification and negation (Socher et al 2013). This means that the system is heavily reliant on a POS tagger and therefore also a degree of grammatical continuity. While this may be a source of error, the general screening of the text suggests that there is a moderate number of grammatical coherent posts.

Using the API access does not give us access to the code but does provide the training dataset which is distinguishes itself by being notably small.

Expectations are that this particular system may suffer from the lack of grammatical context and the small dataset. Though the paper states that they have deliberately moved away from
a bag of words approach (Socher et al 2013) which many other systems use (Pang and Lee 2008), it may be a drawback that the systems is so heavily geared towards the negation through the POS tagger, considering the type of dataset.

7.1.4 Feeling meter

As an extension of the current methods of sentiment analysis, the Feeling Meter was developed in order to be able to detect arousal at the same time as valence (positive/negative). The tool is trained on publicly available Facebook posts where people have tagged their post with one of the 143 available feeling-tags on Facebook (Zimmerman et al., 2015). Before the data was used as a training model, the 143 feelings were bundled together and arousal / valence values were given manually for each bundle. The simple online demo of the tool9 takes a single text, and calculates the level of each of five base feelings chosen by the creators. Values for all the five feelings combined sum up to 1.0 in total. In the current implementation the feeling meter always assumes that feelings are present and therefore it will have difficulties in properly identifying neutral posts (Zimmerman et al., 2015). In the data we received from the feeling meter, the individual posts were all given the values in an Excel sheet.

7.1.5 Semantria

As Semantria is a commercially developed tool, technical specifications of its implementation are not very accessible. For the purpose of this text we have used their blog and our own tests using Semantria as a basis for establishing how it works. Semantria offers both an API access as well as an Excel plugin to use their sentiment engine. The API offers various options for developers and companies to utilize the features of Semantria in their own implementations whereas the Excel plugin makes it easier for one-off analysis as the ones we have done in this paper. For the purpose of this paper they were kind enough to grant us 80000 transactions through the excel plugin. The excel plugin lets the user mark a row of texts (post content) and optionally a row of text ID’s. Once this is selected the data is sent to Semantria and a detailed report is sent back. The report contains a detailed analysis of each text, split into a number of elements. The overall sentiment score for the entire text is given along with a label. Furthermore, the detailed report identifies key entities and themes in the text and gives individual sentiment scores on these. It even gives a Wikipedia category and relevance score for the post.

---

Semantria utilizes a hybrid system applying a mix of supervised and unsupervised machine learning before applying sentiment analysis. The system is divided up in low-level activities (tokenization, PoS tagging), mid (topic identification, intention detection) and high (sentiment analysis) (Redmore 2015)

7.2 Research Design

Various tools differ in bias towards certain data types. We have tried to accommodate such discrepancies by addressing by aligning the approach, tools and data source that neither of the tools are favourably advantaged. In order to outline our consideration, the following sections will explore the choice of datasource, the aim and credibility and the method.

7.2.1 Data source

In order to identify the most reliant datasource, preliminary tests and discussions have been conducted.

The discussions served to exclude data sources early on, such as trustpilot, as we suspected these would not yield a clear result often due to lengthiness. In this process help forums, review sites and discussion forums was discounted.

The initial field was narrowed down to Facebook and Twitter data, 10,000 posts from the same time period with the same topic and object of sentiment were collected from Radian6, from both Twitter and Facebook. The results showed that both had a heavy weighting of neutral posts and thus Facebook was selected, due to a higher polarisation / percentage of negative and positive posts versus neutral. The same results were confirmed via Semantria.

Similarly, Twitter was estimated to have more disruptive elements in regard to #hashtags and @twitterhandles that would favor human annotators far beyond the systems. Facebook was considered to have more coherent, less disruptive data to which neither of the systems would have a distinct benefit.

Radian6 is based on Semantria which is trained on a considerable dataset and is using lexalytics resources. Similarly, The Feeling Meter is trained on facebook data as well but with an overweight of positive posts. The feeling meter is not specialized in strictly sentiment analysis but the valence numbers are extracted for analysis anyway.
7.2.2 Aim, credibility

The experiment is intended to shed light on or confirm assumptions that a) sentiment analysis is inaccurate, by testing prominent tools in the market and b) human concordance is a factor to take into consideration when evaluating the accuracy.

In order to lend credibility to the study, the tools selected carry a certain gravitas. Radian6 is widely used for business purposes as an extension of Salesforce and both Sentistrength and Semantria are well established in the market/field. While it may only decisively conclude on the tools in question and is only for one type of data input, it forms the offset for a larger study comparing the accuracy of sentiment analysis tools.

Similarly, there is also a basis for assuming that human concordance is not perfect (Donkor 2013). We compare the overall agreement between the annotators to decide the degree of agreement between them. A low agreement would potentially invalidate any evaluation of the accuracy of sentiment tools and thus it is a relevant parameter to take into account.

7.2.3 Method

The experiment is divided into numerous smaller tests to ensure accuracy. While the sample size compared between the tool is 18469, a sample of 1000 have been chosen for annotation under the assumption that this random sampling is representative.

1. Collection of data via Radian6
2. Collected data is analyzed via Stanford API, Sentistrength and The Feeling Meter.
3. The different output is normalized for comparison.
4. The overall concordance between the tools is evaluated.
5. A selected sample of 1000 random posts are extracted and annotated with valence.
6. Human concordance is compared and evaluated.
7. Agreement between valence for systems and humans is evaluated overall and on tool by tool basis.
8. Accuracy of the feeling meter is evaluated based on annotations.
9. Interviews are conducted with relevant academics and companies.

7.2.4 Process description

Emphasizing transparency in the process below is a thorough explanation of the process, based on the method described above.
The data for sentiment analysis was collected through Radian6 as it offers powerful tools for collecting public post data from Facebook, Twitter and several other sources. It was also necessary to do this as the sentiment analysis tool of Radian6 only annotates data which it has collected directly. Approximately 50,000 posts were collected but reduced to ~18,500 in order to exclude incomplete posts as Radian6 cuts off posts longer than approx. 300 characters when exporting to csv.

The sentiment from Radian6 was retained and the dataset was later duplicated and the copy had all punctuation replaced with commas creating a dataset labelled WB (with breaks) and NB (no breaks) for the sake of the Stanford classifier.

The two datasets were subsequently run via the Stanford Classifier via their downloaded Java code, Sentistrength (courtesy of Chris Zimmerman), The Feeling Meter (Daniel Hardt) and Semantria via an excel plugin (courtesy of themselves).

Four human annotators, including the authors, have meanwhile annotated the first 1000 posts retrieved by Radian6 in the same order allocating them into four categories: Negative (neg), neutral (neu), positive (pos), unreadable/unrated (unr).

Afterwards, the data was collected, calculated and visualised via a spreadsheet comparing each individual post between human annotators and between human annotators and the systems. Furthermore, a number of aggregate key point indicators were calculated for comparison such as human concordance, aggregate number of positives, neutrals, negatives etc.

These numbers have then been evaluated on in the sections Analysis (9.0) and Discussion (10.0)

Interviews have been conducted with a number of relevant academics and Falcon Social and Trustpilot These have been identified based on prior knowledge and selected based on a preceding discussion of relevance.

The interviews have been informal and only semi-directed. Interviewees have been asked a series of open ended questions with the possibility to further expand on the topics.
7.2.5 Pruning and normalizing text

The preceding treatment and pruning of the text is relatively low. The Stanford Classifier API splits on punctuation, thus prompting us to use two datasets, one including punctuation and one where all exclamation marks, question marks, colons and periods are replaced with commas. Both datasets have been tested in order to establish the impact of this approach.

Initially we collected 50,000 posts via Radian6 but since it would not retrieve the entirety of the posts, all posts ending in “...” have been excluded from the dataset to ensure that all systems had the same dataset, thus same basis for analysis.

Avoiding pruning or normalizing the text to a greater extent, has allowed for a larger study as we have not been forced to manually repair or address text elements text elements. It is considered that impact of significant issues are reduced to a small error margin as the scale of the analysis is increased.

7.2.6 Annotation

The annotation is based on 4 parameters. Negative (NEG) Neutral (NEU) Positive (POS) and unreadable/unratable (UNR). Due to the variation in the systems, the lowest common denominator was chosen. In order to achieve a decent human concordance, it was also hypothesized that fewer options would yield a more consistent annotation between the coders.

Coders were furthermore instructed with an intercoder agreement (Appendix A) to annotate based on the perceived sentiment and not take into consideration knowledge and expectations of the systems capabilities.

They were furthermore instructed to consider the overall sentiment of the sentence not geared towards a specific topic but rather the overall emotion instead of the object of the sentence.

---

If there are several emotions throughout the post, please evaluate the most prevalent one or indicate as neutral, where necessary.

---

Except from Intercoder agreement, appendix A

Page 39 of 101
Annotators were also instructed to disregard political orientation and correctness and to base every annotation on their own opinion *as there are no right answers* and the only time they should seek assistance was in matters of translation.

The annotators have been labelled as C1-C4 with C1 and C3 being respectively the authors of this paper, Peter and Emil. The other two are close acquaintances proficient in English skills, though belonging to a different demographic.

### 7.3 Interview setup

#### 7.3.1 Aim of the interviews

The interviews were carried out in an undirected and exploratory manner. The aim of the interviews are to provide (supporting) qualitative data to reflect on the current approach of the field, in order to clarify potential business applications of the systems which we are addressing.

The secondary aim is to uncover theory relevant to the practices and systems which we are addressing. As computational linguistics encompass everything from computer science to mathematics to psycholinguistics, it is difficult to encompass the entirety both within the paper as well as our literature review. Working with academics specialized across the field allows us to access their knowledge of essential relevant theories. As this is conducted exploratory, this means that the academic interviews have been a formative part of the paper as it would direct us unto other topics which we otherwise would not have encountered. The interviews therefore serve a twofold purpose: identifying relevant theoretical considerations and identifying relevant business applications.

#### 7.3.2 Choice of interviewees.

The processes of identifying interviewees have been based on access to academics in the field as well as our personal knowledge of their work. Attending various workshops and meetings with professionals in the field, we have been granted an insight into their work.

We have identified Maria Klara Stein for an interview due to her work and understanding of feelings in regards to especially Facebook data as well as her understanding of various dimensions of feelings and emotions (Stein 2015). This has been deemed relevant partially in regards to sentiment analysis but also other data and text mining relevancies.
**Chris Zimmerman** was identified through workshops and presentations as well but in particularly in regard to his work with Daniel Hardt on the *Feeling Meter*. As it has been deemed a relevant aspect of sentiment analysis, as well as an indication for potential future uses and business applications, this forms the basis for the interview.

**Dirk Hovy** was identified as relevant also through previous meetings and workshops. Due to his work with both psychographic data and sociographic data to augment the linguistic data. The aspect of determining which data is relevant, and combining datasets rather than relying strictly on semantic and syntactic interpretations, formed the basis for the choice to include him.\(^\text{10}\)

**Falcon Social** has a social media monitoring platform with a sentiment analysis tool integrated. Similarly, they have access to a wide range of social media data and utilize this data in various way. Their work with data, and the prospects that they are looking for ways to further utilize this data, prompted an opportunity for an interview addressing what their intentions are for harvesting insights.

**Trustpilot** has access to a significant database of company specific and rich text data with a tremendous potential for mining insights. In addition, they have access to extralinguistic data via people filling in their bio. With the prospects that they might be interested in utilizing this data, we also organized an interview.

Interviews were scheduled with Infomedia and IBM (Watson), but unfortunately cancelled due to unknown circumstances, at a time that did not allow for rescheduling.

---

\(^\text{10}\) Unfortunately, the recording was lost due to some technical complications, hence we have chosen to refer primarily to his respective papers, and refrain from raising specific points or quotes from the particular interview.
8.0 Results

In the process of analyzing our test data, four human coders annotated the first 1001 posts from the same dataset as was run through the different tools. The coders were all given the same instructions and were all asked to follow the intercoder agreement (Appendix A). Halfway through annotating the posts, coder 4 contacted us to clarify the approach, as he thought he might have misunderstood something. Instead of rating the sentiment in the posts objectively, he had annotated the sentiment as seen from the point of view of the Trump campaign, which was topic of the data collection. Instead of requesting the coder to start over, we asked him to complete annotating using the same approach. This is the same way as the Radian6 tool calculates sentiment and thus we thought it interesting to have data annotated in the same way by a human coder.

8.1 Cohen’s Kappa

To measure the intercoder-agreement on an individual post level, two different methods have presented themselves. First, there is Cohen’s Kappa (Viera et Garrett 2005) which gives an indication of the agreement between only two specific coders. To calculate the agreement across all of the coders, if would be necessary to use Fleiss’ Kappa (Fleiss 1971) which can calculate across several coders. In this case, we have calculated the Fleiss’ Kappa for coders 1-3 and for all the coders to compare the two. Besides that, we have calculated the Cohen’s Kappa for coder 1 and 3, as they seemed to be closest, based on the overall agreement at around 70%, in our results.

<table>
<thead>
<tr>
<th>Fleiss’ Kappa for Coder 1-3</th>
<th>0.403547699035734</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fleiss’ Kappa for Coder 1-4</td>
<td>0.335824734751406</td>
</tr>
</tbody>
</table>
| Cohen’s Kappa for Coder 1&3 | 0.671624713958811

The kappas themselves are not very informative without comparison. It is clear to see that the agreement between coders 1-3 is quite higher than if coder 4 is taken into account. The agreement between coder 1 and 3 is also significantly higher. To put the kappa’s into context the table below can be used. It was first presented by Landis & Koch in 1977 and has been widely used as a benchmark for kappa’s since. However, the benchmark has met some

---

11 Appendix B - Kappa calculations
critique (Gwett 2014) as it does not have a substantial basis and as it can be difficult to use the same benchmark for what is good in cases with everything for two to n coders. In this paper we have chosen to include the Kappa benchmark from Landis & Koch (1977) as it still serves as some sort of indication of what the Kappa scores mean.

<table>
<thead>
<tr>
<th>Interpretation of Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Kappa</strong></td>
</tr>
<tr>
<td>&lt; 0</td>
</tr>
<tr>
<td>0.01–0.20</td>
</tr>
<tr>
<td>0.21–0.40</td>
</tr>
<tr>
<td>0.41–0.60</td>
</tr>
<tr>
<td>0.61–0.80</td>
</tr>
<tr>
<td>0.81–0.99</td>
</tr>
</tbody>
</table>

Figure 7 - As first presented by Landis and Koch (1977), found in: (Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. Fam Med, 37(5), 360-363.)

8.2 Sample representativeness of whole dataset.

The annotated posts are a sample of 1001 of the posts from the full dataset. Therefore it is important that we consider how these systems harmonize with the entire dataset when considering the overall distribution. While it can still be considered a relatively small sample in big data contexts, it is still large enough to be considered statistically significant at around 18.500 posts (there are slight variations between the systems)

<table>
<thead>
<tr>
<th></th>
<th>Breaks</th>
<th>Breaks</th>
<th>Nobreaks</th>
<th>Breaks</th>
<th>Nobreaks</th>
<th>Breaks</th>
<th>Nobreaks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Radian6</td>
<td>Sentristrength</td>
<td>Feeling Meter</td>
<td>Semantria</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEG</td>
<td>12.69%</td>
<td>38.86%</td>
<td>39.26%</td>
<td>19.18%</td>
<td>17.08%</td>
<td>31.07%</td>
<td>31.47%</td>
</tr>
<tr>
<td>NEU</td>
<td>79.42%</td>
<td>39.66%</td>
<td>39.66%</td>
<td>20.08%</td>
<td>19.48%</td>
<td>45.15%</td>
<td>44.76%</td>
</tr>
<tr>
<td>POS</td>
<td>7.39%</td>
<td>21.48%</td>
<td>21.08%</td>
<td>60.74%</td>
<td>63.44%</td>
<td>23.78%</td>
<td>23.78%</td>
</tr>
<tr>
<td>Full</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NEG</td>
<td>NEU</td>
<td>POS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEG</td>
<td>12.54%</td>
<td>37.66%</td>
<td>38.00%</td>
<td>19.65%</td>
<td>18.02%</td>
<td>30.31%</td>
<td>30.17%</td>
</tr>
<tr>
<td>NEU</td>
<td>79.68%</td>
<td>36.87%</td>
<td>37.28%</td>
<td>21.86%</td>
<td>20.05%</td>
<td>44.30%</td>
<td>45.00%</td>
</tr>
<tr>
<td>POS</td>
<td>7.48%</td>
<td>25.48%</td>
<td>24.72%</td>
<td>58.51%</td>
<td>61.95%</td>
<td>25.39%</td>
<td>24.84%</td>
</tr>
</tbody>
</table>

Appendix C - Data results
It becomes clear that the 1001 posts overall are very close to being representative of the full dataset. The highest differences are seen at sentistrength with around 3% fluctuations of some numbers. Overall, the sample can be said to be representative of the whole dataset. Stanford results have been left out of this table - see section 9.2 for an explanation.

8.3 Human Concordance

The human concordance is an estimate of how often the coders agree with each other. Since we have only operated with 4 coders, the value has also been calculated coder-to-coder as well as aggregate. Aggregate is listed as C123 and C1234. The reason for separating C4 out is that the annotation has been carried out with sentiment directed towards “Trump” rather than just evaluating the overall sentiment and is therefore considered as a separate entity.

<table>
<thead>
<tr>
<th></th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>61.34%</td>
<td>71.33%</td>
<td>60.24%</td>
</tr>
<tr>
<td>C2</td>
<td></td>
<td>63.14%</td>
<td>57.34%</td>
</tr>
<tr>
<td>C3</td>
<td></td>
<td></td>
<td>62.94%</td>
</tr>
<tr>
<td>C123</td>
<td></td>
<td></td>
<td>49.35%</td>
</tr>
<tr>
<td>C1234</td>
<td></td>
<td></td>
<td>36.26%</td>
</tr>
</tbody>
</table>

Appendix C - Data Results

The overall concordance between individual coders is relatively low, averaging 62.7%. However, considering the agreement between all coders, it is significantly lower, at only 49.4%. C4 who has annotated different further contributes to a significant drop in overall agreement, reducing it to 36.3%.

“There have been various studies run by various people and companies (including Biz360 before it merged into Attensity), and they concluded that the rate of human concordance is between 70% and 79%.”

*(Donkor, 2014)*
While there are few mentions of studies in regard to human concordance, this can be considered unusually low. Donkor (2014) states that human concordance is approximately 70-80%, but does not mention whether this is between multiple individuals, or whether it is human-human or human-machine.

The human concordance only scores above 70% between coder 1 and coder 2, and thus it should be noted that this happens to refer to the authors of this paper, thus suggesting an inclination towards either a similar mindset or perhaps the influence of the inherent knowledge of the systems.

### 8.4 Human-to-system agreement

<table>
<thead>
<tr>
<th>% Agreement</th>
<th>Radian6</th>
<th>Semantria</th>
<th>Feeling Meter</th>
<th>Sentistrenght</th>
<th>Stanford</th>
</tr>
</thead>
<tbody>
<tr>
<td>WB</td>
<td>WB</td>
<td>NB</td>
<td>WB</td>
<td>NB</td>
<td>WB</td>
</tr>
<tr>
<td>C1</td>
<td>53.65%</td>
<td>50.85%</td>
<td>47.35%</td>
<td>30.17%</td>
<td>29.27%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>46.75%</td>
<td>45.45%</td>
</tr>
<tr>
<td>C2</td>
<td>45.35%</td>
<td>50.15%</td>
<td>47.65%</td>
<td>32.37%</td>
<td>31.17%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>46.55%</td>
<td>47.15%</td>
</tr>
<tr>
<td>C3</td>
<td>55.34%</td>
<td>51.55%</td>
<td>49.15%</td>
<td>28.17%</td>
<td>27.07%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>47.15%</td>
<td>46.65%</td>
</tr>
<tr>
<td>C4</td>
<td>58.14%</td>
<td>42.46%</td>
<td>41.86%</td>
<td>21.78%</td>
<td>20.68%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>39.56%</td>
<td>39.86%</td>
</tr>
<tr>
<td>C1234</td>
<td>53.12%</td>
<td>48.75%</td>
<td>46.50%</td>
<td>28.12%</td>
<td>27.05%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>45.00%</td>
<td>44.78%</td>
</tr>
<tr>
<td>C123</td>
<td>51.45%</td>
<td>50.85%</td>
<td>48.05%</td>
<td>30.24%</td>
<td>29.17%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>46.82%</td>
<td>46.42%</td>
</tr>
</tbody>
</table>

Appendix C - Data Results

Between the coders and the systems, there are various levels of agreement. For each column, the high rate of agreement is marked by a green cell. These numbers are an expression how often coders and systems agree on specific posts not the aggregate number of pos/neu/neg.

While C4 is closest to Radian6 - as per the expectations - it is not significant compared to C1 and C3. However, C4 has notably lower agreement with Semantria and the feeling meter (8-10% pts.) and somewhat compared to Sentistrength (6-7 percentage points.)

The aggregate agreement for Radian6 varies significantly due to C2.
The aggregate agreement for Semantria is remarkably close for C1-C2-C3.
The feeling meter has a low overall agreement but it should be noted that it is modified for the purpose. While it outputs a negative and positive value totalling 1, it has been modified so that 0.33-0.67 is calculated as neutral. It scores equally low, compared to all the dedicated sentiment systems.

Sentistrength has a similar agreement among C1, C2, C3 for both WB and NB and while outperforming The Feeling Meter, it scores slightly lower than Semantria and significantly lower than Radian6.

The difference in C123 and C1234, while not calculated to consider the weighting of only having one different annotator, does still give an indication of how large the fluctuation is through C4’s way of annotating. In all cases, it seems to have increased the overall average.

### 8.5 Sentiment Distribution across systems

<table>
<thead>
<tr>
<th>Total Posts</th>
<th>Tool</th>
<th>WB</th>
<th>NB</th>
</tr>
</thead>
<tbody>
<tr>
<td>18471</td>
<td>Radian6</td>
<td>12.54%</td>
<td>79.68%</td>
</tr>
<tr>
<td>18352</td>
<td>Semantria</td>
<td>30.31%</td>
<td>44.30%</td>
</tr>
<tr>
<td>18352</td>
<td></td>
<td>30.17%</td>
<td>45.00%</td>
</tr>
<tr>
<td>18466</td>
<td>Feeling Meter</td>
<td>19.65%</td>
<td>21.86%</td>
</tr>
<tr>
<td>18466</td>
<td></td>
<td>18.02%</td>
<td>20.05%</td>
</tr>
<tr>
<td>18469</td>
<td>Sentistrength</td>
<td>37.66%</td>
<td>36.87%</td>
</tr>
<tr>
<td>18469</td>
<td></td>
<td>38.00%</td>
<td>37.28%</td>
</tr>
<tr>
<td>18420</td>
<td>Stanford</td>
<td>92.23%</td>
<td>3.29%</td>
</tr>
</tbody>
</table>

*WB: Dataset with breaks (punctuations)*

*NB: Dataset with no breaks (punctuations replaced with commas)*

**Appendix C - Data Results**

The distribution of sentiment varies across systems. The Stanford Classifier has only been able to run a dataset where the breaks or punctuation ( . : ? ! ) have been replaced removed and replaced with commas, while Radian6 which has originally gathered the 18471 posts, has retrieved and analyzed the posts with breaks. Therefore, the data is tested with and
labelled to indicate the numbers for Sentistrength, the Feeling Meter and Semantria using both datasets.

It should also be noted that there are certain technically related discrepancies resulting in a slight difference in the total number of posts, however not enough to significantly alter the aggregate results.

As the numbers show, there are a wide span in the distribution. Most notably is Stanford, which is extremely weighted towards negative sentiment at 92.2%.

While Semantria seems affected by the different dataset (WB and NB), neither the Feeling Meter and Sentistrength appears has a significant alteration to the numbers with a max fluctuation around 3.5%.

### 8.5.1 Semantria Distribution

The output from Semantria spans from -1 to 1, down to 15 decimals, allowing us to visualize the distribution thusly:

![Figure 8 - Distribution of sentiment in Semantria](image)

It should be noted, that the data is cleaned as there are 6000 posts at exactly 0 which have been left out of the visualization. While there appears to be an even distribution across the
numbers, there are various scores that spike significantly - such as the before mentioned 0, and 0.6 at 501 occurrences, 0.49 at 359 occurrences. As visible, there are a number of outliers raising questions to why some values are so overrepresented. However, the overall distribution seems fairly even, though the overall trend leaning towards neutral at around 44%.

Comparing to Sentistrength which scores on a range from -2 to 2 with intervals of 1, we get a similar image, though not as fine grained:

![Figure 9 - Distribution of sentiment in Sentistrength](image)

While neutral (0) seems overrepresented, it is important to consider that -2 to -1 both account for the number of negative. This is not considered in the simplified 3 stage evaluation used by our annotators. However, the nuanced image does inform us of a larger lower degree of very positive or very negative than moderately negative/positive posts, though the levels of negative posts seem to be similarly distributed between moderately and very negative compared to very and moderately positive. This adds depth to the otherwise even distribution when regarding the numbers for Sentistrength.
Figure 10 - Sentiment Distribution of the 4 annotators

<table>
<thead>
<tr>
<th></th>
<th>NEG</th>
<th>NEU</th>
<th>POS</th>
<th>UNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>26.17%</td>
<td>58.04%</td>
<td>15.78%</td>
<td>0</td>
</tr>
<tr>
<td>C2</td>
<td>33.67%</td>
<td>47.55%</td>
<td>18.48%</td>
<td>0.30%</td>
</tr>
<tr>
<td>C3</td>
<td>25.47%</td>
<td>61.84%</td>
<td>12.69%</td>
<td>0.00%</td>
</tr>
<tr>
<td>C4</td>
<td>22.28%</td>
<td>66.83%</td>
<td>7.59%</td>
<td>3.30%</td>
</tr>
</tbody>
</table>

Appendix C - Data Results

A similar image is presented when looking at the human annotation. With no exception, neutral posts are the predominant evaluation. C2 has a significantly more even distribution, whereas C1, C3 and C4 follow a much more similar pattern, though with C4 scoring a lot lower on positive posts. This can likely be attributed to the different way of annotating. Considering the numbers however, C2 is the most notably different in regard to distribution.
### 8.6 Aggregate comparison of humans and systems

While the human concordance is overall fairly low, the aggregate result between humans and systems provides a different measure of accuracy.

<table>
<thead>
<tr>
<th></th>
<th>NEG</th>
<th>NEU</th>
<th>POS</th>
<th>UNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>26.17%</td>
<td>58.04%</td>
<td>15.78%</td>
<td>0.00%</td>
</tr>
<tr>
<td>C2</td>
<td>33.67%</td>
<td>47.55%</td>
<td>18.48%</td>
<td>0.30%</td>
</tr>
<tr>
<td>C3</td>
<td>25.47%</td>
<td>61.84%</td>
<td>12.69%</td>
<td>0.00%</td>
</tr>
<tr>
<td>C4</td>
<td>22.28%</td>
<td>66.83%</td>
<td>7.59%</td>
<td>3.30%</td>
</tr>
<tr>
<td>C123 (AVG)</td>
<td>28.44%</td>
<td>55.81%</td>
<td>15.65%</td>
<td>0.10%</td>
</tr>
<tr>
<td>C1234 (AVG)</td>
<td>26.90%</td>
<td>58.57%</td>
<td>13.64%</td>
<td>0.90%</td>
</tr>
</tbody>
</table>

#### Appendix C - Data Results

<table>
<thead>
<tr>
<th>Total Posts</th>
<th># Tool</th>
<th>NEGATIVE</th>
<th>NEUTRAL</th>
<th>POSITIVE</th>
<th>MIXED</th>
</tr>
</thead>
<tbody>
<tr>
<td>18471</td>
<td>Radian6 WB</td>
<td>12.54%</td>
<td>79.68%</td>
<td>7.48%</td>
<td>0.3%</td>
</tr>
<tr>
<td>18352</td>
<td>Semantria WB</td>
<td>30.31%</td>
<td>44.30%</td>
<td>25.39%</td>
<td>0.00%</td>
</tr>
<tr>
<td>18352</td>
<td>Semantria NB</td>
<td>30.17%</td>
<td>45.00%</td>
<td>24.84%</td>
<td>0.00%</td>
</tr>
<tr>
<td>18466</td>
<td>Feeling Meter WB</td>
<td>19.65%</td>
<td>21.86%</td>
<td>58.51%</td>
<td>0.00%</td>
</tr>
<tr>
<td>18466</td>
<td>Feeling Meter NB</td>
<td>18.02%</td>
<td>20.05%</td>
<td>61.95%</td>
<td>0.00%</td>
</tr>
<tr>
<td>18469</td>
<td>Sentistrength WB</td>
<td>37.66%</td>
<td>36.87%</td>
<td>25.48%</td>
<td>0.00%</td>
</tr>
<tr>
<td>18469</td>
<td>Sentistrength NB</td>
<td>38.00%</td>
<td>37.28%</td>
<td>24.72%</td>
<td>0.00%</td>
</tr>
<tr>
<td>18420</td>
<td>Stanford NB</td>
<td>92.23%</td>
<td>3.29%</td>
<td>4.48%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>
While neither of the systems are notably close to the human annotation, there is still a significant difference in the aggregate distribution compared between human annotators and systems.

The Stanford classifier may be discounted as the numbers are so significantly different than both other systems and human annotators.

The Feeling Meter is likewise leaning towards a very different distribution being very positive-heavy.

Radian6 appears to have a similar development as the human annotators, though with significantly more neutral heavy.

If we consider how many percentage points the different systems vary, it becomes apparent that the systems differ significantly in how close they are to the numbers the human annotators:

Figure 11 - Comparison of distribution for Radian6, Semantria, Feeling Meter and Stanford
<table>
<thead>
<tr>
<th></th>
<th>Dif-C123</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radian6</td>
<td>WB</td>
</tr>
<tr>
<td></td>
<td>NB</td>
</tr>
<tr>
<td>Semantria</td>
<td>WB</td>
</tr>
<tr>
<td></td>
<td>NB</td>
</tr>
<tr>
<td>Feeling Meter</td>
<td>WB</td>
</tr>
<tr>
<td></td>
<td>NB</td>
</tr>
<tr>
<td>Sentistrength</td>
<td>WB</td>
</tr>
<tr>
<td></td>
<td>NB</td>
</tr>
<tr>
<td>Stanford</td>
<td>NB</td>
</tr>
</tbody>
</table>

**Appendix C - Data Results**

These numbers are calculated based on the difference in each category (NEG, NEU, POS) then calculated how much each category differs between human and systems, and then added up, hence allowing for a greater number than 100%-pts.

It appears that the performance of Semantria strongly outperforms the other systems. Sentistrength also appears to have a generally good performance and both systems do not appear particularly affected by the difference in the datasets WB and NB.

**8.6.1 Systems score on individual posts**

As previously mentioned the tests have been done on two datasets, one with breaks (punctuations) and one where the breaks have been removed from the posts. With the data we got from Sentistrength and the Feeling Meter, it was possible to compare the individual posts and how the system had annotated these.

By comparing the two sentiment columns in Excel, and having it put “Match” or “No match” as a result in a third column, we were able to count how many times Sentistrength and the Feeling Meter actually agreed with it’s own sentiment based on the different datasets (with and without breaks).
The results were as follows:

<table>
<thead>
<tr>
<th></th>
<th>Feeling Meter</th>
<th>Sentistrength</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Match</strong></td>
<td>17302</td>
<td>11257</td>
</tr>
<tr>
<td><strong>No match</strong></td>
<td>1167</td>
<td>7212</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>18469</td>
<td>18469</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Match %</th>
<th>No match %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Match %</strong></td>
<td>93.68%</td>
<td>61%</td>
</tr>
<tr>
<td><strong>No match %</strong></td>
<td>6.32%</td>
<td>39%</td>
</tr>
</tbody>
</table>

*Appendix C - Data Results*

(Sentiment matches on the two datasets - with/without breaks)

As we can see in 39% of the posts Sentistrength did not give the same sentiment to a post, even though the only changes in the post were removed punctuations. The Feeling Meter scores significantly higher where it agrees with itself in almost 94% of the posts.

Only the Feeling Meter and Sentistrength are included, since the other systems did not allow for individual comparison due to the way posts were split and numbered (Stanford, Semantria) or the systems were only able to process the one dataset (Radian6).

### 8.7 Interviews

The interviews can be divided into two categories - with academics and with businesses. While there are overlapping topics, generally the focus of the interviews with businesses was founded more in current applications while the academics were more theory based and speculative.

#### 8.7.1 Academic interviews - Key takeaways

The main takeaways from the interviews were in regard to future potential and current applications.

It was argued that sentiment in its current form is ineffective and uninformative (Zimmerman, 2015), even though Falcon use it as a valuable metric.
Future applications are highly tied to the inherent meanings of text, or the semantics. For example, how to detect sarcasm, emotional state or the intention to vote for Trump behind the statement “Trump 2016”. Similarly, it was noted that text data carries a tremendous potential on the semantic level, which is currently not utilized. This is exemplified by topics such as profiling, topic mining and emotional contagion. The challenges hereof are also how systems can identify these elements often based on very limited information/data.

The topic of extra-linguistic data was also raised in relation to psycho and sociolinguistic data. The question of how to implement data about the premises of the text (e.g. what forum) or data about the author (e.g. via a profile on the forum) and the value of this information.

The interview with Mari-Klara addressed a number of different topics, especially in regard to feelings, e.g. how expressed feelings aren’t necessarily an accurate representation. It was also discussed the reason for developing the Feeling Meter and the relevance of feelings in a data analysis context.

### 8.7.2 Business Interviews - Key takeaways

**Falcon Social** offers a social media publishing, managing and monitoring tool. The tool offers sentiment analysis based on data that is analysed through the Semantria API. The monitoring part of the tool allows the users to analyse the sentiment of posts on their social media profiles, but is also able to push alerts to relevant people, if for example the percentage of negative posts suddenly increases. Their experience with sentiment is that it is quite neutral-heavy [Alfredsen, 2015 - #00:02:43-5#], and their results tend to be even more neutral, as they only want to present their customers with a “positive” or “negative” label, if they are really certain it is correct. Looking at individual posts usually does not give the best results, but the aggregate sentiment gives a better indication of what people are saying [#00:02:54-7#].

One of their biggest wishes for future development of their NLP capabilities, is to be less dependant on Semantria but also develop a better form of topic identification and extraction [#00:04:25-2#].

Recently, they started offering a “Facebook topic data” feature where they monitor sentiment towards a specific manually chosen topic, for e.g. features of a new car, to see which features are best received [#00:09:12-4#].

A recurring point made in the interview was that even though NLP offers a lot of new and interesting ways to analyze all of the available data, one of the big challenges with
implementing it at companies like Falcon is that they have to be able to communicate their results in a meaningful way to customers who are not NLP experts [#00:12:14-1#].

**Trustpilot** is using NLP technologies in order to extract useful data from their reviews. As explained during the interview, they have divested from using sentiment analysis due to inaccuracy and instead rely more on the ratings. This corresponds well with our preliminary results, as well as a study mentioned by Dirk Hovy, stating that there seemed to be no relation between the rating and the overall sentiment of Trustpilot posts.

The focus has instead been on using topic mining to identify core topics and concepts discussed in the reviews. The process is still predominantly manual, in order to identify the key influences and topics. Sentiment is used in combination with either very positive or negative posts, in order to identify what is spoken negatively and positively of, and what they say of it. This allows for a specific analysis of what people value, discuss and what the specific issues are, e.g. if customers generally talk about delivery in the context of good reviews. (Alfredsen, 2015)

The process is still very manual relying on a fair amount of work hours and Semantria’s excel plugin. The future aspects are greater automation as well as including further extra-linguistic data such as the demographic data derived from their user profiles.
9.0 Analysis

9.1 Evaluation of sentiment analysis

9.1.1 Cohen’s Kappa and Human concordance - do we agree?

When doing manual annotation (coding), NLP researchers typically calculate a kappa value to give an indication of the agreement between coders. One of the most widely used is Cohen’s kappa which can calculate the agreement between two coders. The kappa is a value between 0 and 1 that can be used to compare the intercoder agreement to other sets of manually annotated data. The issue with Cohen’s kappa is that it only calculates between two coders and in our case this would not be sufficient. As an alternative to Cohen’s kappa, when working with more than two coders, there is the Fleiss’s kappa. Both kappas were calculated in this paper where the Cohen’s was only calculated between the authors.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fleiss’ Kappa for Coder 1-3</td>
<td>0.403547699035734</td>
</tr>
<tr>
<td>Fleiss’ Kappa for Coder 1-4</td>
<td>0.335824734751406</td>
</tr>
<tr>
<td>Cohen’s Kappa for Coder 1&amp;3</td>
<td>0.67162471395881</td>
</tr>
</tbody>
</table>

Appendix B - Kappa Calculations

As mentioned in the results section, the kappas are compared to a benchmark from 1977 by Landis & Koch. The benchmark has met some critique as it does not have a documented basis but is more an expression of the author’s idea of what a good kappa is. Some of the critique is based on the statement that a good kappa is relative to the data, the coders and the entire setting. If we have a hundred random people reaching a kappa of 0.5, this would be very good, but a kappa of 0.5 for two coders that have the same background might not be seen as good.

Looking at the kappas we got from our tests, we see that the agreement between coder one and three (the authors of this paper) is significantly higher than when we take coder two into account as well. As we see coder four, who rated in a different way, significantly lowers the kappa, compared to when it’s only 1-3. Some of the differences between the coders individually will be further discussed later on. The kappas have been included as they are standard in these studies and to be able compare the agreements in this study to other studies later on more easily.
9.1.2 Tendencies within human concordance - *why do we disagree?*

When considering the human concordance, it becomes obvious that annotators overall agree very little on the specific posts. It also becomes clear that there are a number of reasons. Of course, C4 should be widely discounted in this exercise as earlier mentioned.

_I don't know about you Donald Trump, but I know lots of lovely Immigrants, Latin and Mexicans and I've never been raped by any._

Above statement has been rated as both positive, neutral and negative by human annotators. Similarly Radian6 has rated it as positive, Sentistrength and Semantria as negative, and the Feeling Meter as neutral. It is clearly a conflicted statement. However, it also illustrates well why there may be a difference in opinions.

While it is an attack directed at Donald Trump, “I don’t know about you..” , this is embedded in the semantics, so it is likely that the systems will not pick up on this particular aspect. “lots of lovely immigrants” may be considered rather positive but the question arises around “I’ve never been raped by any”. The statement is not positive, when considered separately. In the entire context, it may not be the focal point of the statement, and a human annotator may weigh the influence on the sentiment low.

The statement adds a negative connotation to the sentence, both when considered separately and in context. This may be the reason for annotating it as negative, if this is considered the prevalent feeling. On the contrary, the coder annotating it as positive may have granted the first statement more weight.

Such statement, and similar mixed-emotion statements, create frequent disagreements. Mixed feeling statements are common where two sentences - often a post including a quotation and a relation to the quote or two differing statements - have a different sentiment. Such evaluation may yield a large fluctuation in what a post is evaluated as. One might consider to demand a specific approach within the intercoder agreement, where the process of annotating would be confined within stricter guidelines, such as annotating each sentence in a post, then calculating the aggregate score, similar to Sentistrength. Sentistrength scores each sentence with a number from -2 to 2 and then takes the average. Such a process among human annotators may create a less subjective and emotional evaluation of the posts. It would, however, disregard the aforementioned aspect of discounting or disregarding
some statements on either sides of breaks where they can be considered less influential on the actual sentiment of the sentence.

Reviewing the agreement among the individual posts, the human annotation appears to be low not because annotators utilize the entire span between neg-neu-pos on many posts, but rather that the posts are predominantly annotated as neutral and one other feeling.

It appears that in most cases disagreement is between neutral and negative or positive. Using a three-stage model does not provide significant insight. As illustrated by the sentiment distribution of Semantria, a wider consideration may be more informative. A more fine grained consideration could likely provide more even results, as there is no clear “line” for when human annotators switch from neutral to either positive or negative. In practical terms, asking annotators to allocate a numeric value, would perhaps be inconsistent and particularly labor intensive, but either using a slider or a wider range of options, could lessen the burden, while providing more insights. For example, Radian6 utilizes both the idea of very negative or very positive as well as mixed, albeit only rarely. While the initial assumption was that a more complex annotation would yield a lower concordance, a too simplified version, such as this, appears to make correct annotation of gray areas more inaccurate. Striking a balance between simplicity for the annotator and useful data, it is perhaps worthwhile considering an alternative to standard labels and instead try to apply values based on a slider. However, utilizing a wide range of values does not address the aforementioned aspect of mixed-emotion statements.

9.1.3 Systems and humans - not always in agreement

When looking at the the match-rate between coders and systems on an individual post level, we see that many coders and systems average around the 50%. The highest agreement we see is between coder 4 and Radian6 which makes sense as coder 4 annotated with sentiment towards Donald Trump, just as Radian6 did.

We generally see that the coders mostly agree with Radian6 and Semantria, even though a 50% agreement is not very impressive. In the discussion later on we will argue that this individual agreement level might be a result of the sentiment categories lacking nuances, which then makes the results less precise.

One thing we see in the results is that the average agreement with Radian6 is higher (53.12% vs. 51.45%) when we include coder 4 in the average. This is different from all the
other systems, where including coder 4 means a lower agreement average. This emphasizes that the "sentiment towards" approach in Radian6 appears to do what it claims, as it - in theory - annotates in the same way as coder 4 does.

9.1.4 Different sentiment distribution across systems

Since the agreement between the systems and coders on individual posts is not very impressive, we have also compared the aggregate results. In this section, we look at the distribution of posts rated in the different categories investigating the hypothesis that even though systems and humans might disagree on an individual level, the aggregate numbers might total to the same. In our tests we see that the aggregate distribution for neutral posts is almost 80% for Radian6 which is a lot higher than our average human aggregate for neutral posts, at 58.57%. This shows that even though the coders agree with Radian6 on more than 50% of the individual posts (one of the highest coder-to-system agreements), the overall result of the dataset varies to a large extent:

![Aggregate distribution of sentiment across systems](image)

Figure 12 - Aggregate distribution of sentiment across systems

One of the most eye-catching things in this graph is the distribution of negative posts in the Stanford tool and neutral posts in Radian6. This is such a significant variation from the other datasets that we later on in this paper discuss why the data from Stanford should not be considered in these comparisons (Section 9.2). It also raises the question of whether the overall agreement on individual posts for Radian6 may be because of the high number of neutral posts simply matching all the neutral posts as annotated by human coders. At least
the individual agreement may be high while the aggregate number is far off the average sentiment distribution, as annotated by human coders (Section 8.3). We also see that the Feeling Meter is heavily weighted towards positive sentiment which is also discussed separately (Section 9.4). Looking closer at Semantria and Sentistrength, we see that the aggregate numbers are fairly close to each other. Semantria is slightly more biased towards neutral where Sentistrength is a bit more negative-heavy, but the numbers are still quite close to each other. We also see that the aggregate distribution of sentiment is almost identical on the tools, whether using the dataset with breaks or the one without. When we compared how often Sentistrength agreed with itself on the individual post level, we saw that it was only in 39% of the posts. Based on this, it appears that the breaks in the dataset have a limited impact on most systems, except Stanford.

9.1.5 System architecture - how certain systems are geared towards certain data

It becomes clear throughout the results that the systems are very different weighted towards certain sentiment. Whereas Stanford is extremely prone to classifying statements towards negative, the Feeling Meter tends towards the positive.

The explanation may lie in which data it is trained on as well as applied to. The Stanford API is trained on a very small dataset, albeit unclear what it is derived from, it clearly does not perform well here.

We see a similar case with the Feeling Meter. The results are overall very positive weighted. While it is trained on a balanced set, it is derived from a very positive heavy set. Considering the distribution of the sentiment and the posts, it would appear that the system is very weighted towards high valence (sentiment), though somewhat more evenly distributed in regards to salience (arousal).
Since the Feeling Meter is trained on Facebook data, it does not yield an accurate result in this aspect. One could speculate on the nature of the posts between this and the training dataset. The training dataset has only posts associated with a certain feeling, since it is derived from feeling tags, which might suggest a low prevalence of profoundly objective posts. While the constructed training dataset is a somewhat equal distribution of various degrees of sentiment, feelings such as ‘meh’ are underrepresented in comparison to very positive tags (Zimmerman et al 2015).

*Figure 13 - Valence and salience of the Feeling Meter*
Arnold Schwarzenegger has been hired to replace Donald Trump as the host of The Celebrity Apprentice. Terminator takes Trump's place on Apprentice] which is scored 0.98 valence and 0.34 salience.

While the salience appears reasonable, such a post does not appear to be particular positive. Considering the training dataset, one might, in this particular case, suspect that the sequence *(has) been hired* will be very heavily associated with posts in the training dataset, due to the nature of the dataset. This will subsequently be associated with highly positive posts.

However, Comparing the two sentences “I was hired” and “I wasn’t hired” only yields a very small difference in sentiment

![Figure 14 - Feeling Meter analysis of the sentence “I was hired”](image)

![Figure 15 - Feeling Meter analysis of the sentence “I wasn’t hired”](image)

This highlights the issue of data disparity as either is presumably associated with a positive or negative feeling. Albeit only moderately positive, ~2.8, it is still significantly more prevalent than fearful and angry (negative feelings).
9.2 The Stanford Classifier and why the data must be discounted

The implementation of the no-breaks dataset was initially intended to allow us to test the Stanford Classifier alongside the other systems. However, certain technicalities would create extra rows and issues and it was deemed to labor intensive to clean the data sufficiently, to allow for comparison of individual posts.

Considering the NB (No-Breaks) dataset, the Stanford Classifier achieved a ratio of negative posts of over 93%, why it has been chosen to disregard this data as too influenced by the removal of the breaks.

The WB dataset was later run in order to consider the aggregate data. Since the classifier splits on punctuation, the number of lines is drastically increased, and does not consider posts holistically, but it might be worth considering if it achieves a high aggregate accuracy.

<table>
<thead>
<tr>
<th></th>
<th>Stanford</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>18420</td>
<td>NB</td>
<td>92.23%</td>
<td>3.29%</td>
<td>4.48%</td>
<td>0.00%</td>
</tr>
<tr>
<td>36100</td>
<td>WB</td>
<td>62.07%</td>
<td>28.40%</td>
<td>9.53%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

Appendix C - Data Results

However, it turns out that it is still very negative-heavy. The accuracy is surprisingly low, compared to ad-hoc tests with their online tool. Reasons are ascribed to the online tool not splitting on punctuation, and thus considering the entirety of sentences. The performance in terms of evaluation of the sentiment on an aggregate level is very poor. Neither does it compare to other systems or human annotators. While the issues with the NB dataset can be ascribed to the influence on the POS tagger, the WB dataset is still extremely negative heavy, according to Stanford. As the output is overly shifted towards the left, it may be negated by altering the thresholds for when something is considered negative. Unfortunately the output is not numeral but only based on very negative to very positive, thus not allowing us to “tamper” with the data, and we cannot confirm whether there is an even distribution of datapoints across the entire range or whether it is very weighted towards the extremes.
We estimate that the poor performance is due to poor grammar and a high presence of punctuation in the dataset, thus fragmenting it significantly which is a disadvantage to Stanford’s bottom up recursive neural model, which is highly reliant on grammatical consistency (Socher et al 2013). Similarly, the training dataset for the system is very small and does not appear to be derived from social data at a quick glance.

The inaccuracy and accessibility of the Stanford Data therefore leads us to discontinue any considerations of this tool due the sub par performance.

### 9.3 Radian6 and coder 4

As mentioned earlier C4 and Radian6 seem to have the highest match rate on individual posts at 58%. Similarly, the aggregate results are also within a relatively small margin of difference.

<table>
<thead>
<tr>
<th></th>
<th>NEG</th>
<th>NEU</th>
<th>POS</th>
<th>MIX</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4</td>
<td>22.28%</td>
<td>66.83%</td>
<td>7.59%</td>
<td>3.30% (Unrated)</td>
</tr>
<tr>
<td>Radian6</td>
<td>12.54%</td>
<td>79.68%</td>
<td>7.48%</td>
<td>0.30% (Mixed)</td>
</tr>
</tbody>
</table>

Appendix C - Data Results

Meaning that the aggregate result has a difference on approx 260 (26%) posts, while the individual disagrees on 420 posts (42%). This is a significant difference, and somewhat curious.
As mentioned earlier, C4 has coded in a similar fashion to Radian6’s annotation which is directed towards a topic (Trump in this case). Some of the discrepancies may be contributed to limitations of the system, but it seems unlikely that this accounts for 42%. It may be that Radian6 does not recognize “Donald Trump” as a hierarchical derivative of “Trump”. It is worth noting that the number of positives are very similar. While this seems to speak to Radian6’s advantage in regard to accuracy, a closer study reveals that there are only 7 of the ~76 positive posts\textsuperscript{12}, that are the same. The aggregate number are curiously close, but the individual posts do not match at all.

Delving even further in, it becomes apparent that this is also the case with the negative posts. Only in 36 instances do the negative posts match, compared to the ~125 negative posts. The high accuracy when considering the individual posts are therefore derived from the high prevalence of neutral posts.

This again raises and/or illustrates the question of aggregate and individual results. And whether one is more accurate than the other.

**9.3.1 Do other systems perform better than R6 on POS/NEG concordance?**

Considering agreement between systems and humans on individual posts, we see that only approximately 50% of the posts for positive and negative, are agreed upon by the humans and the systems. In order words only around 50% of those annotated as positive or negative by human coders, are similarly annotated by the systems.

Seeing as Radian6 performs surprisingly poor in regard to annotation on the individual level, it is worth considering whether other systems perform better with the annotators. The table below shows how many of the individual positive and negative posts the systems and the coders agree on.

E.g. for C1 the person annotated 262 posts as negative, and Sentistrength agree in case of 147 of these posts.

<table>
<thead>
<tr>
<th></th>
<th>NEG</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>147 (262)</td>
</tr>
</tbody>
</table>

\textsuperscript{12} The number of posts is derived from the lowest percentage.
Appendix C - Data Results

It becomes obvious that there are a significant disparity between which posts are actually annotated as positive-negative. Considering the aggregate results suggest that the NEG/POS posts average 27/14 percent for human annotators, 30/25 percent for Semantria and 38/28 percent for Sentistrength, it is surprising that the numbers for individual posts are so significantly different between humans and systems. It would suggest, that while still inaccurate, the aggregate score performs significantly closer to the coders, compared to the comparison of the individual posts.

9.4 The Feeling Meter - Biased and positive-heavy datasets and the influence on results.

We know the Feeling Meter is trained on an “artificially” constructed dataset in order to get a varied dataset between the feelings, though the original collected dataset was predominantly positive heavy. There is a question of a certain bias as the output data seems very positive heavy compared to other systems and human annotators. Our knowledge of the systems allows us to ascribe this to the dataset. However, it could be argued that we would be able to alter the interval which is considered neutral (currently intervals between 0.33-0.67), to be more neutral weighted. It is possible to alter the interval to span a wider range, e.g. 0.1-0.9 (must always sum 1) and this would shift the graph to be neutral, but also reduce both the number of negative and positive posts, while the issue of the system being very positive weighted remains. It could still be argued that since there is no “fixed” interval for when a post can be considered neutral, it is acceptable to alter the intervals. However, the data is similarly deprived of validity in this process as well as the information. It should rather be prioritized that the system is utilized for either it’s intended purpose or the training dataset should be leveled out. Based on the figure in 9.1.5 however, it appears that there is not an
even distribution of posts, and most posts are placed with a valence score between 0.85 and 1.

In order to consider whether the intervals should be adjusted, it is necessary to consider the overall distribution between the intervals. As visible below, the system has a high presence of very positive posts, thus tuning the intervals would not affect the outcome greatly.

Below is the current distribution as well as the distribution with the interval 0.1-0.9 depicted and the distribution of posts on intervals of 0.01 between 0 and 1 positivity.

![Figure 17 - Comparison of average of human annotation and feeling meter](Image)
Figure 18 - Number of Negative, Neutral and Positive posts if interval is shifted to 0.1 - 0.9 for the Feeling Meter. We tested different intervals, to see if it harmonized better with human annotators.

Figure 19 - Distribution of sentiment for the Feeling Meter. Shows why shifting the neutral-interval does not affect the outcome particularly.
The key takeaways are that the systems vary significantly, likely due to their structure and especially due to their training sets. It also shows, that the distribution is not just a question of moving the intervals to span more than the middle third of the values to adjust for the number of neutral posts, to match the human annotators better.

9.5 Semantria - accuracy, intervals and spikes

While some tools, like the Stanford tool, only gives a sentiment label as a result, others also gives us the actual sentiment score as a value. This is the case with Semantria and Sentistrength, which gives us a numerical value from a negative to a positive number. This numerical value is then used to assign the sentiment label. As previously shown in the results, we have created a graph illustrating the distribution of values in the data received from Semantria.

![Figure 20 - Distribution of sentiment values across Semantria, -1 to 1](image)

In the graph, only a single datapoint dot was omitted, in order to make the graph readable. The dot was representing the value 0,0 and had over 6000 occurrences. This goes to show, that even though this graph displays a much more nuanced picture of the sentiment distribution in this graph, compared to labelling with positive/negative/neutral, we still have a
majority of our posts tagged with a value that does not really provide any idea of the content - meaning that a post with the value 0,0 provides absolutely no insight.

Looking in more detail at the data, we see that one of the posts that scored 0,0 in Semantria was the following:

```
So Arnold Schwarzenegger's new catchphrase will be "You're Fired" instead of "I'll Be Back"? Schwarzenegger will replace Donald Trump Arnold Schwarzenegger will be the new host of The Celebrity Apprentice reality show, replacing Donald Trump
```

Why this post in particular was difficult for Semantria to understand, we do not know. We can see in the data that it was able to identify people like Donald Trump and Arnold Schwarzenegger but never calculate a sentiment towards any of them. Perhaps it is due to the very small amount of latent words in the posts but regardless, it could be argued that it should always be possible to find a small change in the values - especially with 8 decimals, as Semantria uses.

### 9.6 Sources of error, considerations

#### 9.6.1 Choice of media

The reason for choosing Facebook data, was an assumption that the snippets would not be too long to yield mixed results, but still produce a clear sentiment, and that they were long enough to contain a sensible syntactic structure. While the latter seemed correct, Facebook posts often yielded sentences with very unclear overall sentiment, due to e.g. quotes or shares in the same sentence.

Similarly, we observed other aspects of this “noisy data” such as a high presence of identical posts, due to shares, a high prevalence of news articles, poor grammar and syntactic structure, namely in relation to qoutes and shares. A number of these are explored in the following sections.
9.6.1 Syntactical and semantic complexities in annotation
While determining the sentiment of a simple statement such as “I greatly disfavour Donald Trump for the presidency” is a straightforward task, the manner of expressing oneself on Facebook often neglects rules of grammar or sentence structure. Similarly, there are certain text bits with little syntactic depth but they can semantically be interpreted as a mildly positive statement: “Trump 2016”. Such a statement could be considered vaguely positive with low polarity, but one could argue that it still displays a high degree of animation.

Throughout the annotation process it became clear that there are certain elements that may influence both how sentiment analysis systems annotate certain posts as well as raise questions and uncertainty for human annotators. Some are, of course, addressed in the intercoder agreement, it does, however, illustrate how certain aspects are not a question of being either positive or negative.

9.6.2 Poor grammar, complex syntactic structure

Republican presidential candidate Donald Trump called the United States 'a dumping ground for the rest of the world' as he rallied thousands of Texas supporters behind his fiery candidacy and promised Republican leaders he's just getting started. Trump: U.S. 'a dumping ground' for the world

Poor grammar or complex syntactic structures may influence the outcome of the evaluation. While humans are more capable of disregarding such issues, machines are prone to error in this aspect. The sentence above illustrates a strongly animated, negative stated sentence: “a dumping ground for the rest of the world” along side with a strongly animated, positive sentence “… as he rallied thousands of Texas supporters behind his fiery candidacy…”. The complex syntactic structure makes for a difficult sentence to decipher as the determiner of the sentiment contains elements that would suggest that the sentence is definitely not neutral. While it may be mixed feelings, there are no doubt that it contains a high degree of animation and/or intensity. While Radian6 does contain a “mixed” feeling category, it very rarely classifies any posts within this segment.

While the sentence may suggest a negative sentiment, it does not infer a negative opinion about republican presidential candidate Donald Trump.

Fox Business "We have a president that doesn't know what the hell he's doing." Addressing a crowd of over 20,000 people in Dallas, Texas, Donald J. Trump called President Obama "incompetent" and promised that no one could do a better job as president than him.
This also illustrates an issue where a POS tagger may struggle to understand, who *him* is referring to.

#dumptrump #nutjob #scaryracist  
Trump: *"It's disgusting what's happening to our country"*  
DALLAS (AP) ? Republican presidential candidate Donald Trump is renewing his campaign against illegal immigration, *telling a cheering crowd of thousands in Dallas* that *"It's disgusting what's happening to our country."*

The example above illustrates a post that has been shared with the addition of user contributed hashtags. Because of the hashtags the overall sentiment of the post is negative, but the hashtag form makes is difficult for machines to pick up on this.

### 9.6.3 Irony/Sarcasm, jokes

Donald Trump becoming President of the United States would be like Jar Jar Binks becoming Senator of Naboo

The above sentence invokes a positive feeling, or has a positive connotation, but does not have any particular positive identifiers. While being neutral syntactically, semantically it is positive - or negative if the object of analysis of Donald Trump.

However, the question is raised in regard to annotation, how should jokes be classified? Are they also a subject to opinion? I.e. a Donald Trump supporting annotator may not find the joke particularly amusing and thus considering it somewhat negative. Other people may agree that it is positive. However, classifying this is a question beyond arousal and valence.

### 9.6.4 Opinion

It is stressed in the intercoder agreement that opinion should not interfere in the evaluation of the sentiment. It should still be considered whether it might present an issue. Stating that "climate change is a hoax" is a somewhat plain statement but may swing to either positive or negative based on whether the annotator considers either as commonly accepted. Some

---

13 Author's notes: In the Star Wars universe, Jar Jar Binks becomes senator of Naboo and passes a vote granting emergency powers to Senator Palpatine, effectively laying the foundation for the formation of the first galactic empire.
may see it as an attack on commonly accepted (and proven) science, while others find that it reinforces their assumptions.

“I’m voting for Donald trump” is of course negative, but the sentiment can be construed as a type of positive reinforcement of the person’s intended actions, thus being positive.

It did not appear to be an issue among the annotators, based on post-evaluation of their answers, but it is still a consideration and worth addressing in the intercoder agreement.

However, it becomes obvious in a few cases where an author of a post prompts the annotator to consider the opinion presented and the feeling intended to invoke:

Trump 2016, you mad bro?

This sentence utilizes the positive reinforcement as mentioned earlier, in “Trump 2016” but also emphasizes a negative aspect or invokes a negative feeling in regard to the context insinuated. The post may be very confrontational overall and could thus be construed as negative. While it does not contain any particular negative words or formulations, it does invoke an association of a generally controversial or negative situation surrounding it. It may be positive towards Trump, though a human would annotate it as negative. A system would likely annotate is as neutral due to the lack of syntactic context.

Despite a short sentence, this also illustrates the limitations of the semantic understanding of sentences, as well as disparity of data as it has a very little syntactic base.

9.6.5 Object of analysis

“I love coke, but pepsi is the worst”

The statement above is great if you are Coca Cola, not as much if you are Pepsi. It raises the question of considering sentiment analysis not on the basis of the entire post - or even just the sentence - but by extracting the the specific elements which are addressed in the sentence, similar to Radian6.
The output of any system would likely be random if simply analyzing the entirety of the post. However, other elements may be lost if you only consider part of a post as the wider context may also have an impact on the overall sentiment. Sentistrength and Radian6 do address this in various ways. Sentistrength analyses the sentences as individual entities and then calculates the total sum:

The text 'I love coke, but pepsi is the worst' has positive strength 3 and negative strength -3

**Approximate classification rationale:** I love[3] coke, but pepsi is the worst[-3]
[sentence: 3,-3] [result: max + and - of any sentence][overall result = largest of posSentenceTotal, negSentenceTotal] (Detect Sentiment)

Radian6 attempts to relate the sentiment towards a certain topic. Testing various selected topics, there are indications that it has an influence on the aggregate sum of positive and negative posts. The system may be set to extract all posts about coke and analyse the sentiment towards “coke”, thus not considering the latter part of the sentence relating to Pepsi.

This also means that the annotation is slightly different for the systems as they consider sentiment in a different manner. However, the large prevalence of neutral posts and a comparison of the posts, which are clearly positive or negative of Trump, are still marked as neutral.

What we see in a tool like Semantria is the ability to identify and evaluate separate elements of the text. Even though an average score for all parts of the sentence is given as the overall sentiment, it is still possible to see more nuances in the post by looking at the individual results from Semantria.

**9.6.6 Context, continuity and emotional contagion.**

One hypothetical question was raised: “How does the state of mind affect the annotation decision” and further “Would certain genres of music affect the annotation decision?”. The questions were derived from the discovery that identical posts appearing throughout the document would occasionally differ in the annotation. It was theorised that this could stem from a degree of emotional contagion, the general state of mind that the annotator is influenced towards.
Characteristic for many of the posts in question is the degree of polarity or the issue of a mixed signal sentence, meaning that it is not immediately clear what the sentiment may be. Often it was characterized by a number of the previously mentioned issues.

Since the post is a news article and the opinion is relatively objective, the sentiment may be stated as negative due to the prevalence of the words critical, criticize, criti- (a stemmer may consider them identical), which is negatively laden. This refers back to the previous issue that the statement is not necessarily negative in opinion, however the prevalence of negative laden words may leave the annotator with a negative feeling as well as systems relying on a simple word-bank.

9.6.7 Shelf life of words, domain adaptation

One prevalent issue in the development information extraction and sentiment analysis is the use of domain specific expressions and ambiguity of words. Domain specific words require specific training corpuses for sentiment tools, POS-taggers, including such words. This is especially true within the medical industry using numerous acronyms and numerous medical terms which are not found in typical texts.

A word such as wicked may predominantly be used as an expression for terrible, evil - in a context as the wicked witch. However, in some areas or between a certain group of people, it may be used with a positive connotation “The work you did on this thesis was wicked”. While this might be difficult to decode on a lexical basis, applying additional extra-linguistic information may help give such an indication. For example, a positive connotation of the word may be constricted to use within e.g. Boston or among a certain age group (or even more specific a specific personality profile coupled with age and location). This can give clues, or indication that a particular use of the word has a different meaning. On a larger scale, geographical or linguistic background could be relevant for distinguishing different meanings between American and British English as well as generally different meanings ascribed to words.
This is likely to affect systems relying on an annotated word-bank as human annotators should to be able to distinguish the meaning by the context. The posts extracted are not geographically limited, thus it does not seem like a prevalent issue in the tested dataset.\textsuperscript{14}

\textbf{9.6.8 Prior knowledge of systems}

Though the study is relatively small, in relation to the number of human annotators, it serves as an attempt to highlight the issues of human concordance. Throughout the process, it has become obvious that this issue merits a study of its own.

One issue of discussion is the background. One aspect is cultural background - e.g. Danish people are more inclined to rate more moderately than americans on trustpilot (Johannsen et. al., 2015). The same aspect may be the case for sentiment annotations as one could suspect certain backgrounds to have different attitude towards certain ways of using language.

Understanding of the language is also an issue as certain idioms or more delicate unspoken elements of the language is lost on people who are not fluent in the language in question. In this particular case, we have only used 4 annotators including ourselves. Despite everybody being considerably fluent in English, there is nonetheless a difference in level.

In addition, we have strived to disregard our own knowledge of the systems as it may influence our annotation significantly from any “ordinary” annotators. For example, we widely acknowledge that the systems do not possess the ability to identify sarcasm as thus we would be prone to annotate it based on the meaning of the written words. We have strived to abstract from this as it was deemed to adapt the interpretation of data favourably towards the systems in relation to testing the real-world application of the systems.

\textsuperscript{14} Author’s note: The credit for these ideas should be ascribed to Dirk Hovy of CST, as he mentioned them in our interview. However we are not able to quote this interview, as the recording was lost.
9.7 Interviews

9.7.1 Emotional arrays, valence and salience - the question of two-dimensional sentiment analysis

Sentiment analysis works with just one dimension, a positive-negative scale or valence. However, as recognized by the paper Emergence of Things Felt: Harnessing the Semantic Space of Facebook Feeling Tags (Zimmerman et al 2015), and subsequently the Feeling Meter has argued that the Salience dimension is overlooked. This parameter adds another dimension to the sentiment analysis, arousal. This stems from a dimensional perspective assuming that feelings such as sad may be negative but low arousal, while anger may still be negative but high arousal. (Mari-Klara interview)

“So from low arousal—calm, relaxed, tired—to very high arousal like excited. Excited is both high arousal and positive. So, for example, agitated or angry or -- is also quite high arousal.”

(Stein, 2015 [#0:07:56#])

This dimensional approach fits well with the idea, already used in computational linguistics, where sentiment is a linear representation.

In addition, this representation applies well with numerous of the earlier mentioned problems as regards the human annotation such as

TrumpSpeak - coming to an elementary school near you .... Donald Trump Speaks Like A Third Grader ? Republican Base Laps It Up Donald Trump knows his audience.

While some may recognize this as positive is does not carry anything of specifically positive. However, human annotators may likely agree that this is highly aroused and all systems
have annotated this as neutral. The Feeling Meter accurately annotates this as very high arousal (0.96).

While this does address the issue of posts which do not carry an obvious feeling of valence, the relevancy of this is still a question. Of course, together they may represent a more specific feeling thus being able to pinpoint a feeling such as sad or angry. This particular post also illustrates how something which is mocking Trump, but is not negative or positive in an observable way. It may be considered positive if the intention is to provoke a laugh, or negative if one is considering the overall support for Donald Trump, raising the issue of the analyst's question to the data. However, this is, as mentioned, a good example of something which has high arousal albeit not a clear sentiment. The implications are is still unknown, albeit slightly more informative than a post merely labelled as neutral.

So does this extra dimension add more information? For the current applications, the added value of another parameter may still rely on sentiment (valence) as the driving component.

![Figure 21 - Russell, J. A. 1983. “Pancultural aspects of the human conceptual organization of emotions,” Journal of Personality and Social Psychology (45:6), pp. 1281 – 1288.](image)
If we consider the dimensional approach, similar to Russell (1983), such a visualisation of the distribution of sentiment on e.g. social data, similar to the way Falcon Social applies sentiment analysis, it may give some insights in the reaction. Similarly, it may be possible to analyse on how a piece of content is received and shared on Facebook (and possibly other media). For example, content or ads promoting users to adopt a sponsor child may seek to invoke other feelings than a barbie advertisement. As social media monitoring goes, this may also be an option to see if certain feelings are invoked in regard to certain mentions of the brand.

While Russell’s (1983) representation may not be perfectly accurate, it does visualize the concept of arousal and valence quite well.

As mentioned in the interview with Mari-Klara; considering valence may also be relevant when developing content aimed at going viral as viral content are often high arousal (Stein 2015). However, it should be mentioned that a large degree of the inherent value of such systems is the automation and application on a greater scale rather than analysing individual posts as this is usually done better by human annotators.

It was also brought up that perception of risk depends on the mood. It can perhaps be speculated into, in connection with conversions on websites, if it is possible to induce a certain mood in order to make potential customers more likely to add to cart or book a call.

9.7.2 Mood, Emotions and feelings - multiple opinions
Emotions and feeling are quite different things but there are also consideration into mood as well as social rules, especially in high-context cultures, for how and what is expressed. Considering the data in the Zimmerman et. al. paper (2015), it is clear that mostly positive posts are the generally most expressed ones. The system may therefore be applied well to the Facebook domain but other rules may apply for other data sources such as conversations and perhaps even other social media. Assuming that the training data applies to all Facebook data is indirectly also an assumption that the manner in which people express themselves when tagging a feeling, is the same manner in which they will express when not tagging it. In different cultural contexts this may not be the case. Underpinning this with the data, it seems that people are more keen to tag positive feelings. However, they all work in the domain of public posts, so when considering e.g. comments or generally public posts, it may be correct to assume that people generally are aware that they are expressing
themselves in front of (possibly thousands of) strangers thus abide by similar social rules, in regards to what is socially acceptable to express in public.

A feeling may be tied to a certain post but it may not be representative for the general mood. One may share a sad post trying to invoke sympathy with the people involved, without being in a particular sad mood. One could argue that frequent and consistent posts of a similar feeling may constitute a certain insight into the mood of the person. Similarly, one should also consider a feeling different from emotions. While the complexity of this reaches beyond the scope of this particular paper, prevalent and relevant examples may be posts containing multiple opinions and a specific object of the invoked feelings - appraisal theory. (Stein 2015 [#0:07:56#])

I like Trump and his Ump but please Donald don't say "I HAVE SO MUCH MONEY" again !!!

While this may not constitute the difference between feelings and emotions, it still illustrates the concept of an underlying opinion on the topic, while the overall expressed sentiment can be considered as negative.

It also ads a different perspective to the previously raised issue of posts which include more, possibly contradictory statements. While some statements may be an expression of a feeling, other may be an expression of emotion or context. In some instances, such as the one above, it may be possible to identify both a negative and positive sentiment. Many sentiment systems rate towards one specific, overall feeling of the statement but this is a perfect illustration of why this form of sentiment analysis may not be a viable, functioning approach for deriving proper insights. In the cases where posts contain contradictory statements it may not be possible to simply consider the post on a one- or two dimensional approach. A post containing a very outspoken positive and negative post, even when considering salience, cannot be classified as a particular emotion. It may be possible to argue that positive and negative statements outweigh each other and thus the sum should be considered, or perhaps that neutral statements with high arousal are rarely neutral, but instead maybe more worth considering what emotions are invoked throughout the post, maybe even in regard to certain topics.
10 Discussion

10.1 Is the current level of sentiment analysis a viable, sustainable and informative approach from a business perspective?

10.1.1 Is sentiment sufficiently accurate?

Though our study is small, it appears clear that the accuracy can be considered low overall. Not only between humans and systems, but also humans in between.

Firstly, addressing human concordance, it should be noted that it varies significantly between coders, and we will later discuss how the technical form can be improved. Considering the concordance, the highest being around 70%, it can be said to be overall low. It can be argued that a sentiment analysis system with an accuracy of 70% would therefore be highly accurate, though it can also be considered adding to the margin of error. What we also observed in the analysis is that neither between humans and systems, nor in between humans, is the concordance very high on specific posts, but the aggregate level seems to be more accurate. While Semantria and Sentistrength outperform the other systems notably, the confidence level is still fairly low especially, in regards to individual posts where they only agree on a maximum of 50%. Considering such a remarkably high margin of error, it raises the question whether the systems give a misrepresentation of reality. Turning an eye towards the distribution, we can see that the average number of positively (26.9%) and negatively (13.64%) annotated posts differ significantly from Semantria (30.3% and 25.4%) which is the most accurate system in this case. This representation is still appropriate in regard to which sentiment has the highest presence, but is clearly not accurate enough to give an accurate representation of the actual distribution (according to the annotators). However, in regard to business application, it may still have a working purpose as the most common role in business is likely social media monitoring, which after all is more occupied with a sudden change. If it is used as a metric or benchmark to evaluate how positive or how negative the conversation is, the inaccuracy of distribution and the high number of neutral posts complicates the use of the systems.
10.1.2 Common pitfalls - where are the systems ineffective?

“I love coke, but pepsi is the worst” is a neutral statement according to Sentistrength. In human eyes, whether this is negative or neutral, it depends on whether you are Coke or Pepsi. An objective observer may find that this is, indeed, neutral as each clause seems to cancel each other out. It is just one of the examples of the issue of contradiction in posts and how the current systems are often unable to deal with it. The dataset provided numerous multi-sentence posts including both positive and negative statements. Some issues may be contradictory of the topic - a post including another post or statement and subsequently the author expressing his or hers own opinion on the post or the matter.

A hard lesson, but keeps repeating, not to mess with Donald Trump. Trump is a decent man and well-respected businessman and he’ll make the best President ever in this country. BOOM: Stephen Colbert Compares Trump to KKK, Gets BRUTAL Surprise 24 Hours Later He had this coming to him.

As the example above shows, there may be two different statements within the same post. One positive towards Trump and one negative towards Colbert. Systems such as Semantria and Radian6 allow the user to select what topic the sentiment should be towards in case of multi topic posts. However, Radian6 does not seem to perform very well on this matter. This is similar the case with Trustpilot where Brian Linde (Linde, 2015) discards the idea of using Semantria for identifying relation or opinion towards certain topics in the wider dataset because it is simply too inaccurate.

In extension, the question of poor grammar also plays in:

How many black American knew of this Donald Trump as president this would be the norm in the United States of America You gone learn today The first 911 attack in America
This has about as much grammatical correctness as *furiously sleep green ideas colorless* (Chomsky, 2002) and is clearly a negative post but shows why POS taggers would struggle with some of the Facebook posts.

The matter of dealing with this is addressed through some systems to an extent. Both Semantria and Sentistrength are to some extent geared towards social media data, even the Feeling Meter is trained on Facebook data, however it is still difficult for systems to accurately evaluate text that does not follow grammatical rules, unless there are a prevalence of non-negated strongly positive and negative words.

While this raises a question of challenges to the syntactic analysis, there is also the question of semantic challenges such as irony and suggested meanings. E.g. a statement such as “Trump 2016” may have no negative words but still have an overall positive connotation. Irony or sarcasm also occurs throughout the dataset but the semantic nature of it makes it impossible for most systems to detect.

It is pitfalls such as, irony, poor grammar and contradictory statements that make out many of the pitfalls which overall affects the effectiveness of sentiment analysis in this particular analysis. We see that the systems are challenged in a number of areas but these seem to be the major issues in regard to achieving a systematic, structured, accurate method of sentiment analysis, which can be applied across different datasets.

**10.1.3 Falcon Social & Trustpilot - Benefits and criticism.**

Sentiment analysis is being implemented in more and more businesses as we see from our interviews with both Falcon Social and Trustpilot. The tool offered by Falcon resembles Radian6 in some ways because it mainly focuses on the monitoring of social media data. Tools like these still utilize the simplest form of sentiment analysis by simply labeling the posts acquired and presenting the data in dashboards. We found in our tests that when presented aggregate, these sentiment score might actually be able to give some idea of the distribution of sentiment in a large dataset. However, the aggregate sentiment of a large dataset does not give much valuable insight for a company - in fact the most mentioned feature which uses sentiment analysis is the ability to alert social media managers when a “shitstorm” is starting and there is a larger-than-usual amount of negative posts made. Interestingly enough, the uses of NLP in the businesses we have looked at (and consider market leaders) seem very limited to sentiment analysis in its simplest form - with labeling. Both companies seem interested in topic identification where Trustpilot has by far come furthest in regards of generating business from it. However, their implementation still seems
to be at a very early pilot-project state where reports and analysis are done manually in a one-of manner and only offered to specific clients. In our interviews with Trustpilot in particular, we found that they seemed reluctant towards investing in implementation and automation of many of the technologies they were working with. According to them, their clients did not seem to be interested in the data. This does not bode very well for the future of NLP applications in business if companies mainly look at short-term capitalization on things that their clients request. In our opinion, the opportunities in NLP are still so undefined and not very commonly known that data companies need to lead the way and show their customers what is possible.

The fact that both Trustpilot and Falcon utilize the Semantria API, and that Radian6 is built on the Lexalytics engine also utilized by Semantria, shows us that the supply of good and reliable sentiment tools still seems to be lacking some competition.

10.1.4 Is the current technical form informative?

10.1.4.1 Objects, subjects and entities

In our evaluation of Semantria, as well as in some of our market interviews, we saw that an interest in a form of topic identification is on the rise. The interest very likely comes from the fact that Semantria offers these features and that both Falcon Social and Trustpilot use Semantria to obtain sentiment data. We saw from our interview with Trustpilot, that they actually use the topic identification possibilities in Semantria more than the actual sentiment score. According to Trustpilot, their customers were able to get more valuable insights from linking the topics identified with the actual Trustpilot score instead of the sentiment score, which in their opinion was not always very precise.

The implementation at Trustpilot is very interesting as it illustrates how a company, which might be considered industry leaders when it comes to data analysis, still struggles to apply sentiment scores in a meaningful and valuable way. The fact that topic identification and extraction is such a big focus for both Falcon Social and Trustpilot is extremely interesting to the authors of this paper. Both companies indicated in their interviews that they were looking into the possibilities of topic identification in order to add extra value to sentiment scores, an idea which the authors of this paper suggested and outlined in a paper a year ago (Hilborg & Nygaard, 2014). Whereas sentiment in itself can be insufficient for explaining data, the addition of topics in the form of entities in the text can give a more nuanced picture of the data. If a post is both negative and positive, the addition of topics or entities could explain what the sentiment is towards. In user reviews, like at Trustpilot, it can be used to identify customer satisfaction with various parts of their experience from only a short text and a star
rating. Radian6 does have its own take on the sentiment towards an entity approach but in their model, the user has to specify the entity manually. What Semantria allows, and what could be interesting to most companies with large amounts of social data, is the automatic identification of topics across the dataset and sentiment towards these topics.

10.1.4.2 Sliders, dimensions and concordance.

In section 9.1.2 we consider the question of rating the posts based on a slider rather than 3 groups of negative to positive. Before starting the human annotations, it was discussed how to build the experiment and the intercoder agreement, to ensure the most reliable human concordance, without setting too many rules that would artificially raise the overall concordance (e.g. disregarding irony). It was initially argued that limiting the number of options would ensure an overall agreement as annotators wouldn’t be in doubt. For example including “very” such as “very negative” could have created more options and therefore a larger scale to disagree on. Of course, it could have been negated to be compared as the base feeling (all “very negative” being converted to just “negative”). Throughout the experiment, it became clear that the overall human concordance was remarkably low. While it was approximately 70% between C1 and C3 on individual posts, the numbers generally indicated a significant disagreement. Then considering Semantria and The Feeling Meter’s output on a numerical scale, it raised the issue of polarity.

Assuming that we operate with a scale from 0 - 1, where values 0.33 - 0.67 may be considered neutral, a 3-step system would consider two posts differently, though they may only vary by 0.1 sentiment score. A post scoring 0.67 may be considered neutral whereas a post at 0.68 would be considered positive. This contributes to a lower overall concordance, also between humans. It was observed that posts are rarely annotated with both negative and positive but often negative and neutral or positive and neutral. This emphasizes the issue of considering the distribution on a 3 step scale. It was therefore determined that future annotations ought to include a slider rather than the 3 step model. This would still allow annotators to indicate the overall sentiment of feelings but would make it possible to visualise a more appropriate distribution through a scatterplot. This may perhaps give a more accurate representation of agreement and evaluation of sentiment both for systems and human annotators.

Another consideration is when including a secondary variable, such as the arousal in addition to valence, when considering the Feeling Meter. While it does hold the potential to be useful and informative, the question is the degree of instruction necessary for the annotators. Implementing two sliders may be effective, though one could also argue that
they are difficult to compare individually. The usefulness in regards to business application or general analysis is somewhat limited, though both Falcon Social and Trustpilot expressed an interest in the possibilities. It is somewhat reliant on sub-plotted areas for certain feelings in coordinate system to properly understand the output as arousal is still a relatively new term. Understanding what feelings typically tie to certain areas may be useful as it would initially be reduced to a relatively simplistic consideration. It may be informative as an extension of a sentiment tools knowing whether it is angry or sad as it contributes to a very interesting aspect of both the analysis and the discussion. There are of course some accuracy issues as depicted in the results at least when it comes down to valence. Unfortunately, arousal is somewhat difficult to interpret without a clear sentiment. This is emphasized with the general issue of many of the systems being neutral heavy. The argument could of course be to simply disregard the arousal in this aspect, though it raises the issue of perhaps mixed feelings and contradiction in the post. While a post can be using both positive and negative statements, and thus scoring overall neutral, it may also score high or very low arousal. This could be an indication of it not being neutral but perhaps being both negative and positive meanwhile being difficult to determine that it is mixed feelings throughout the post. The current shape of the system, however, makes it difficult to extract such information from each posts as they seem to evaluate on an aggregate level (of the post) similar to Sentistrength. This is also generally a challenge to the system as a POS tagger is difficult to implement properly for social data, which often doesn’t contain correct grammar and spelling. Without the syntactic analysis, it may be more difficult to appropriately break down the sentence structure and thus derive different clauses and inherent topics and objects.

There is no question that an additional dimension may provide additional and in time valuable insights, however, the application and accuracy are still an issue.

Finally, it should be noted that human concordance cannot be measured as the number of times annotators agree on individual posts since a larger group of annotators would disagree more frequently. Adopting a slider and implementing floating values would provide a better output.

**10.1.4.3 Boatloads of neutrality - how do we filter the data?**

As observed in the datasets there is a high presence of repeated posts and news media, which does not add to the conversation. Throughout the results we saw high percentage of
neutral posts. The human annotators cited an average of 58.5% of the dataset as neutral. The systems vary between 20 and 79.7% indicating the disagreement on the proportion.

This raises the question; how do we filter the data effectively? While shares may give an indication of gravitas of a certain post and can therefore be considered as an expression for an opinion, other posts such as news articles do not add to the conversation. For example, 89 of the 1001 posts (~9%) are news articles containing the term Arnold Schwarzenegger. In this instance it may be acceptable as it adds a neutral dimension to the dataset, however it also shows that the relevancy of the posts vary a lot based on the objective. If we assume that the intention with sentiment analysis is to identify the general sentiment of a conversation regarding a certain topic, then should news outlets not be excluded?

Throughout the annotation we noted that many of the news articles were characterized by being repeated frequently. One solution may be to disregard or remove posts that occur several times or have a very high similarity to other posts which are also generally neutral.

Many of the repeated posts were about Arnold Schwarzenegger replacing Donald Trump on Celebrity Apprentice. Semantria may be able to identify a high frequency of Schwarzenegger and Celebrity Apprentice, indicating them as topics, giving an insight into the general topics of the conversation. Due to the repeatedness of the posts, it could be argued that the systems retrieving the data should have a threshold for when posts are typically similar to a range of other posts, excluding these in the aggregate score as they will tend to be news articles. This may exclude other often shared posts which is, of course, also a consideration. On the other hand, they may be relevant to analysing the conversation based on what one is looking for.

10.1.4.4 Why the value of sentiment is overstated - Our results / Topic oriented and Emotion array oriented sentiment analysis

From our experiments with the various sentiment analysis tools and interviews with relevant people in the industry and researchers, we have seen that the current applications and implementations do not offer a very valuable product. One of the main criticisms is that the labeling of sentiment does not give a well-nuanced representation of the actual data. For manual annotation, many of the disagreements between the coders were ascribed to the fact that the coder was in doubt whether the text was neutral or positive/negative. For manual annotation, our suggestion is to use a slider that assigns a numerical value instead of choosing labels. Labeling is also inaccurate when looking at sentences with multiple
opinions or sarcastic parts. Instead, an implementation of topic identification as seen partly in Semantria could give much more insight into what is actually said in large datasets.

Another issue is the performance of current sentiment analysis. As most tools are trained on manually annotated data, and since people rarely agree, a system which performs similar to (or better than) humans is especially hard to devise. The training data also makes the system biased as we saw in the Feeling Meter example in particular where an over representation of positive tagged posts might have been the cause of the system being positive-heavy. The fact that the training data is very influential is also backed by the fact that Semantria has recently started offering specialty packages for restaurants and hotels.

Lastly, the current performance of sentiment analysis on the individual post level is far from usable. The only way that we actually see uses of sentiment data is when the aggregate is analysed. This gives an idea of what the general sentiment is, but not much more.

We have previously explored if sentiment could be made more informative if directed at certain topics. In a previous paper “Towards a more informative approach to sentiment analysis”, we discussed a preliminary process model based on Brill (1997):

Figure 22 - Brill’s 3-step process model (Brill, 1997)

In this process model we suggest applying a 3 step model for sentiment analysis. Firstly, determining the topic, identifying all sentences with a given keyword. Secondly, applying sentiment analysis to these clusters of sentences. Thirdly, using a POS tagger or an unsupervised clustering method to identify which (e.g.) adjectives are commonly associated with specific topics.
Throughout our data collection we were presented with a similar approach at Trustpilot, albeit sufficiently inaccurate to fully automate, indicating that it requires significant human interaction to ensure accuracy. Considering issues such as hypo- and hypernymy, fully automated topic identification does pose some challenges as some terms may have multiple variants. In addition, these are the challenges we also see Falcon Social address (Alfredsen 2015) in order to extract more insightful information on an aggregate level. Instead of considering each individual post, they are working on developing a system which can intelligently identify and give insights on what the conversation is, evolves around and if there are changes to the conversation. Considering the inaccuracy of the sentiment, which we have seen in our tests, we suggest augmenting sentiment analysis with a degree of information extraction, such as topic mining, to derive insights on the actual topics of the conversation. Further being able to extract actual opinions and movements within the conversation, poses significant challenges to the current level of systems as they must be able to aggregate and sum up the “movements” or the general points of the conversation. While Trustpilot manages to derive some insights into satisfaction on different parameters, helped by the review ratings and Semantria, Falcon Social’s project to develop a system capable of identifying topics before classifying and categorizing opinions is an ambitious project.

Relating back to our initial suggestion on how to construct a system, we have previously experimented with unsupervised clustering. One particular issue, which is also likely to present itself for both Falcon Social and Trustpilot, and others alike, is the question of data disparity. In order to derive valid insights, there needs to be a significant amount of data.

10.2 What are the future areas for NLP and sentiment analysis?

The internet has seen more and more rich text data being generated. One such observation is how more and more interaction is becoming text-to-text interaction such as the emergence of discussion forums or using Facebook pages for customer service inquiries. The following suggestions to future applications, are based on our own discussions and speculations, based on the research done in this paper, and the interviews conducted with relevant people actors in the field.

10.2.1 Extralinguistic data - Profiles and contexts

NLP technologies have augmentation potential from extralinguistic data. While there are some sociolinguistic and psycholinguistic aspects which may be inferred, in order to gain insights related to profiling, this is still at an early stage. More simplistic extra linguistic data
may, however, generate some insights in various contexts. For example communication patterns and expression may vary between geographic background, demographic data or context.

Expressing oneself in a tech-help forum may be significantly different from the type of language used on a comment on a friends photo or an opinion piece in a newspaper. This general context should be taken into account when deriving insights from the text. Understanding how humans express themselves in different contexts may be relevant to increase accuracy and develop systems capable of transcending different domains.

Geographical and demographic data can provide similar insights. As described in (Johannsen 2015), there is a difference between age groups and genders when reviewing, when looking at how positive and negative they generally rate things, although the reviews themselves tend to remain indistinguishable in sentiment according to systems.

Additional information in connection with demographics etc. may also generate insights in the manner in which certain age-groups express themselves. Languages tend to be dynamic, with the meaning of words and expressions changing, and new words entering the language. Some of these are geographically determined such as the difference between American and British English. As previously mentioned, a word such as “wicked” may for most age groups be associated with a negative connotation, whereas others may consider this extremely positive “That was a wicked party last night”. One benefit is to appropriately understand the sentence and derive the correct meaning. Another is deductively deriving insights about the author - if one word is particular to a certain group of people or area it may be possible, within a degree of probability, to determine that the author is likely part of this group thus deriving possibly relevant insights.

10.2.2 Profiling - Personality Insights from text

In addition to adding the extra linguistic data, data from social media can also help identify personalities of the authors in a much more specific scope. In a recent paper, Plank & Hovy (2015) have identified and mapped 1500 people who on Twitter identified with a personality type from the popular Myers-Briggs test. Around 1.2 million tweets from these people were then gathered and used as a training model for an NLP tool. The tool was then found to be able to identify certain aspects of the Myers-Briggs personality traits, only from reading tweets from a subject. The paper argues that this is currently one of the only research projects with this much data to analyze, but also that the data is biased as they only use
Twitter as a source. In spite of biases, and a still limited number of subjects (1500), we still see this paper as a clear example of what NLP can be used for. Using models like this one, extra linguistic profiling can be greatly enhanced and easily automated as it is possible to make mechanisms that can automatically add personality information to a profile.

10.2.3 Temporal, geospatial reasoning - determining intentions
In section 5.8 semantic role labelling or shallow semantic labelling are mentioned as a technology. Throughout the work with various tools and interviews, it has been observed in a very limited degree. Only Semantria mentions that it is part of their plan for developing the tool, but no actual implementation has so far been observed. It has a tremendous potential, however, especially in regards to temporal, geospatial reasoning and also for identifying intentions and topics.

10.2.4 Customer service
When looking into the future of NLP, it seems very obvious to look at some of the current projects being done. What Falcon Social and Trustpilot are both doing, are very straightforward applications for especially customer service departments. Trustpilot in itself is very based on customer satisfaction and therefore their opportunities in customer satisfaction and service are quite big. With the current implementations it is possible for a company to easily identify parts of their service which can be improved. With a more automated approach and use of NLP, it could be easier to continuously monitor and track satisfaction with specific services. Combined with the earlier mentioned profiling of the author of a post, there is also an opportunity to help customer service departments evaluate the urgency of a comment, review or e-mail. Profiling the user, combined with an analysis of their language and what they are actually saying, can help identify people who needs an answer now. The urgency can be based on both the customers need for urgent help with the product but could also be based on the likeliness of the post going viral. This way, companies can have a chance of preventing, or preparing for so-called “shitstorms”.

10.2.5 Comment moderation and the Asshole Meter
The Feeling Meter provides us with an interesting secondary metric in the form of salience (arousal). This may prove a relevant metric in regards to very biased or highly aroused comments, comments such as racial slurs. It could also be comments in debates which are only meant to stir up anger or controversy. If a system was to be trained on a biased dataset,
characterized by high-arousal statements, of a tertium non datur type, it could be applied to a degree of comment moderation. One method may be through requiring very high arousal or determined statements to be approved before posted. While it is not relevant for the entirety of open discussion forums, it does have a potential for promoting a proper discourse. Arousal may not be the best indicator as highly aroused statements may just be positive or negative while still containing serious content. However, training on a dataset such as the comment section from Nationen may still provide the system with capabilities of identifying similar patterns of expression. In extension, a system which could encourage more moderate discourse - the aptly named Asshole Meter - could provide a warning based on the arousal, sentiment and general patterns of expression before a toxic comment is posted - urging the user to reconsider a more appropriate or moderated response.

10.2.6 Intelligence

It is safe to assume that various intelligence agencies are implementing NLP technologies in their research and screening. While it (unsurprisingly) has not been possible to get any comments or interviews from people involved or directly related. However, the use is undeniable in regards to screening the amounts of text content being exchanged and generated. Software which is able to monitor certain people perhaps for expressed opinions on certain topics, and possibly through SRL intentions, may offer a degree of efficiency in generating leads for screening. For example, school shootings in the US which has been announced on reddit, or people generally expressing positive opinions about certain organisations or acts, may generate leads of people for surveillance. It should be noted that there are considerable ethical and legal debate around such an extensive surveillance.

10.2.7 Cookie enrichment

Cookie enrichment refers to the process of storing information about a web user as a cookie on the computer. It is mainly used within digital marketing for banner advertising and is often the heart of third party data trading. Companies selling or enriching data often seek to categorize users based on what types of sites they visit in order to identify demographic, interests, family status etc. While some information may be derived from simply a filled in bio or profile on a webpage, other systems seek to identify the topic of particular sites to classify a user as interested in a certain topic.
This presents an opportunity for NLP in the future. A visit to a particular site, e.g. a car site, may indicate that a person has browsed family cars from which you can derive a fair amount of insights. However, a system capable of extracting deeper information from the site or from sites containing rich text, could possibly have enormous potential in regards to automating it.

Firstly, extracting deeper information than current systems may reveal that a site with the keyword “Justin Bieber” may in fact be a “I hate Justin Bieber” site. Developing a easy-to-implement system across websites, which is able to mine and enrich a cookie from the text contained on the site, in relation to sentiment, intention and topics would make this technology more accurate.

In addition, applying a similar technology to sites dedicated to user producing rich text such as value networks (Stabell & Fjeldstad 1998) - Reddit, Trustpilot, disqus, etc. - to enrich cookies in connection with interest and attitude towards different topics. A user posting a question to ask.com about which car is better, may be a prime target to identify for car-retailers. User reviewing the performance of their newly bought car may be relevant as a negative target group for retailers, but a prime segment for insurance brokers. Identifying elements such as attitudes towards topics, and intention of users, may contain tremendous added value to DMP’s and data-trading.

Finally, as an extension of the segment about extralinguistic data, it may also be possible through language use to draw certain conclusions about authors of statements - e.g. if they use a word specific for a certain age group - thus estimating the age group or location of a person/cookie.

10.2.8 Detecting complex dependencies in rich text

As the amount of data generated increases every day, NLP continuously becomes more and more important in order to utilize this data in a valuable way. With machine learning, NLP can both learn from it (by using training models) but also help improve it. When big data is analyzed, NLP can be used to help find patterns and tendencies in the text - something that is not possible by just looking at numbers. One example here could be in medical journals, which are all written in natural language. A specialized NLP model could in this case be used to read and analyze all the journals of a specific group of patients to see if there are any commonalities or patterns in e.g. blood type, reaction to medication, symptoms and so on.
11 Ethics

There are a number of ethical considerations in this technology many of which relate to A.I. and surveillance. In addition, there are certain implications in regards to the collection of data. While the posts collected may be public post, it is probably safe to assume that many people are not aware of the potential application of their data.

Firstly, in relation to collecting data, there are some considerations when it comes to sensitive data. Using Facebook data where each post has an ID allows us to identify the person who posted it. Since the posts are public, one could in theory argue that it is no problem, however it does allow for some misuse of the data. Systems like Google and Facebook ensure anonymity in their targeting systems for advertising (Advertising - Privacy & Terms - Google, n.d.), ensuring that it is not possible to target or identify specific people. The data in this particular paper allows this, but is treated discreetly. ID’s for the posts are retained, but mostly when it comes to ensure uniqueness and being able to sort through the posts. It could be argued that the ID’s could be replaced with a self-generated ID, but it also compromises the validity of the study, as there is no way to confirm the data. However, in this particular case, it may be irrelevant.

The concept of data collection without expressed consent is an issue and it must be assumed that not all people are aware that their post may be considered public domain and object for such analysis. However, since this is not intended towards identifying or being used in a personal context, and the data is treated anonymous and aggregate, we do not consider there being any breaches in regards to ethical behaviour.

With surveillance, there are considerations when it comes to the possibility of a computationally driven mass surveillance. There are often debates or rulings to whether mass surveillance is illegal/unconstitutional (Stempel, 2015) as well as ethical acceptable. Systems are well on their path to being able to identify intentions and opinions allowing intelligence and surveillance agencies to broadly and automatically screen people. The future application of such technologies should therefore be scrutinized when it comes to data laws and general ethics for individual applications.
12 Conclusion

12.1 Sentiment analysis is inaccurate and uninformative

Although we only compare a limited number of tools in this study, we see a clear tendency that sentiment analysis is not very accurate. On an individual post level, the precision seems almost random. Only when looking at an aggregate level of sentiment, the tested sentiment tools delivered a somewhat acceptable output. We see that sentiment analysis is very dependent on the data used to train the model. This gives a very variable quality of results depending on the analysed data. At the same time, training sets often come from humanely annotated data, which raises the issue that a system, trained on data from people who are never more than 80% in agreement, cannot be more accurate.

12.2 Sentiment is not informative on its own

The inaccuracy of the sentiment analysis tools is not only based on the technologies or the training data. From our research we can conclude that the current form of sentiment in itself holds some of the fault for the inaccuracy. Labelling sentiment with most oftenly three different labels (sometimes five), positive, negative and neutral, removes the nuances in the dataset. At the same time, the definition intervals in a model can greatly shift the overall sentiment from one of the outer values to neutral.

We have seen that just defining a text as either positive, negative or neutral is not very informative. At Trustpilot they had started to ignore this sentiment approach as it did not provide value to their customers. However, when combined with extra linguistic data or topic identification, the value rapidly increases and the prospective business application broadens.

12.3 Future applications - which technologies are essential for a working system

We have observed a high variety in the annotation of individual posts and the accuracy can sometimes be ascribed to the domain being social data. Since this data tend to be void of grammatical rules and correct spelling, accuracy of systems relying on lemmatizers, stemmers and POS taggers not designed for these is likely to be inaccurate. We have previously tested unsupervised machine learning, clustering, for identification and while it provided a degree of information, it seemed that a POS tagger could provide more accurate information to the manner in which the words were related to each other. Future applications are therefore likely reliant on hybrid systems, such as Semantria, that relies on unsupervised
machine learning but still applies a series of predefined grammatical rules and lexicalized
text sequences. It is likely that future systems will utilize hybrid approaches in order to
maximize the data collection with the ability to transcend different domains. Development of
wordbanks such as WordNet, VerbNet, FrameNet which infer more insight and
understanding of text, is an excellent indicator that rules and context still have a relevance in
information extraction.

12.4 What industries/fields will it be seen in
As the world becomes more digitized and data is becoming cheaper to gather and store, we
see that all industries can benefit from automatically analyzing data. Among our suggestions
we see intelligence agencies, news media and in general any industry that does customer
support. Sentiment analysis, and NLP in general, can be applied to almost anything that has
unstructured data. We can especially see this in the applications of IBM Watson which,
according to IBM, can be applied to any industry/field. The Watson model is still very
dependant on human annotation and curation but gives a good indication of the broad
possibilities for the technologies.
13 References


Donkor, B. (2014, November 23). SENTIMENT ANALYSIS: WHY IT’S NEVER 100% ACCURATE.


Fuglsang, L. & Olsen, P. B. Videnskabsteori i samfundsviendskaberne, 2004


Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003) - Find subtopics


Predicting iPhone sales from tweets - Lassen, Madsen, Vatrapu 2014


Appendix A: Intercoder agreement

Intercoder Agreement - Evaluating sentiment of Facebook posts.

In order to ensure the quality of annotation, it is important that all annotators abide by a set of basic rules. These rules will provide common ground to ensure that all participants understand the task at hand, and will be able to solve it in a similar manner.

The data which we ask you to annotate will be forwarded as a Google Spreadsheet. Column B will contain the posts, which we ask you to annotate. Note, that there are one post per row. Column C-D-E will be labelled with Negative (C) - Neutral (D) and Positive (E). After evaluating the sentiment of a post, please mark a single x only in the corresponding cell (e.g. a positive post in row 10, should have an x in cell E10)

Also, please note the following guidelines:

When evaluating sentiment, the sentiment should be based on the overall sentiment of the post, not towards a specific topic.

If there are several emotions throughout the post, please evaluate the most prevalent one or indicate as neutral, where necessary.

Annotators should not consider political correctness, opinions or similar, but only evaluate on the expressed sentiment in the sentence.

Post that are incomprehensible, should be tagged in the corresponding column F

The evaluation are individual, and the entire series should be completed without interference from other participants, except in cases of translation.
### Appendix B - Cohen’s Kappa

<table>
<thead>
<tr>
<th></th>
<th>POS</th>
<th>NEG</th>
<th>NEU</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>82</td>
<td>10</td>
<td>35</td>
<td>127</td>
</tr>
<tr>
<td>NEG</td>
<td>13</td>
<td>164</td>
<td>78</td>
<td>255</td>
</tr>
<tr>
<td>NEU</td>
<td>63</td>
<td>88</td>
<td>468</td>
<td>619</td>
</tr>
<tr>
<td>Total</td>
<td>158</td>
<td>262</td>
<td>581</td>
<td>1001</td>
</tr>
</tbody>
</table>

**Agreement**

<table>
<thead>
<tr>
<th></th>
<th>POS</th>
<th>NEG</th>
<th>NEU</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>82</td>
<td>164</td>
<td>468</td>
<td>714</td>
</tr>
<tr>
<td>NEG</td>
<td>20.04595</td>
<td>33.24076</td>
<td>73.71329</td>
<td>127</td>
</tr>
<tr>
<td>NEU</td>
<td>0.671625</td>
<td>0.671625</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>1001</td>
</tr>
</tbody>
</table>

**Kappa**

0.671625 0.671625
**Human Annotators**

<table>
<thead>
<tr>
<th>Total # Posts</th>
<th>Annotator ID</th>
<th>NEG</th>
<th>NEU</th>
<th>POS</th>
<th>UNRATED</th>
</tr>
</thead>
<tbody>
<tr>
<td>1001</td>
<td>C1</td>
<td>26.17%</td>
<td>58.04%</td>
<td>15.78%</td>
<td>0</td>
</tr>
<tr>
<td>1001</td>
<td>C2</td>
<td>33.67%</td>
<td>47.55%</td>
<td>18.48%</td>
<td>0.30%</td>
</tr>
<tr>
<td>1001</td>
<td>C3</td>
<td>25.47%</td>
<td>61.84%</td>
<td>12.69%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1001</td>
<td>C4</td>
<td>22.28%</td>
<td>66.83%</td>
<td>7.59%</td>
<td>3.30%</td>
</tr>
<tr>
<td>Average C1234</td>
<td></td>
<td>26.00%</td>
<td>58.57%</td>
<td>13.64%</td>
<td>0.90%</td>
</tr>
<tr>
<td>Average C123</td>
<td></td>
<td>27.14%</td>
<td>58.74%</td>
<td>12.92%</td>
<td>1.20%</td>
</tr>
</tbody>
</table>

**Agreement between annotators**

<table>
<thead>
<tr>
<th>C2</th>
<th>C3</th>
<th>C4</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>61.34%</td>
<td>71.33%</td>
</tr>
<tr>
<td>C2</td>
<td>63.14%</td>
<td>57.34%</td>
</tr>
<tr>
<td>C3</td>
<td>62.94%</td>
<td></td>
</tr>
<tr>
<td>C1234</td>
<td>49.55%</td>
<td></td>
</tr>
<tr>
<td>C1234</td>
<td>36.26%</td>
<td></td>
</tr>
</tbody>
</table>

**Human - Aggregate System Agreement**

<table>
<thead>
<tr>
<th>Radian6</th>
<th>Semantria</th>
<th>Feeling Meter</th>
<th>Sentistrength</th>
<th>Stanford</th>
</tr>
</thead>
<tbody>
<tr>
<td>WB</td>
<td>WB</td>
<td>WB</td>
<td>WB</td>
<td>NB</td>
</tr>
<tr>
<td>C1</td>
<td>53.65%</td>
<td>50.86%</td>
<td>47.35%</td>
<td>30.17%</td>
</tr>
<tr>
<td>C1</td>
<td>45.35%</td>
<td>50.15%</td>
<td>47.65%</td>
<td>32.37%</td>
</tr>
<tr>
<td>C3</td>
<td>55.34%</td>
<td>51.55%</td>
<td>49.15%</td>
<td>28.17%</td>
</tr>
<tr>
<td>C4</td>
<td>58.14%</td>
<td>42.46%</td>
<td>41.96%</td>
<td>21.78%</td>
</tr>
<tr>
<td>Average C1234</td>
<td>53.12%</td>
<td>48.75%</td>
<td>46.90%</td>
<td>28.12%</td>
</tr>
<tr>
<td>Average C123</td>
<td>51.45%</td>
<td>50.85%</td>
<td>48.05%</td>
<td>30.24%</td>
</tr>
</tbody>
</table>

**Sample vs Full set distribution**

<table>
<thead>
<tr>
<th>Radian6</th>
<th>Semantria</th>
<th>Feeling Meter</th>
<th>Sentistrength</th>
<th>Stanford</th>
</tr>
</thead>
<tbody>
<tr>
<td>WB</td>
<td>WB</td>
<td>WB</td>
<td>WB</td>
<td>NB</td>
</tr>
<tr>
<td>Sample</td>
<td>NEG</td>
<td>12.69%</td>
<td>31.07%</td>
<td>31.47%</td>
</tr>
<tr>
<td></td>
<td>NEU</td>
<td>79.42%</td>
<td>45.15%</td>
<td>44.76%</td>
</tr>
<tr>
<td></td>
<td>POS</td>
<td>7.95%</td>
<td>23.78%</td>
<td>23.78%</td>
</tr>
<tr>
<td>Full</td>
<td>NEG</td>
<td>12.54%</td>
<td>30.31%</td>
<td>30.17%</td>
</tr>
<tr>
<td></td>
<td>NEU</td>
<td>79.68%</td>
<td>44.30%</td>
<td>45.00%</td>
</tr>
<tr>
<td></td>
<td>POS</td>
<td>7.48%</td>
<td>25.39%</td>
<td>24.64%</td>
</tr>
<tr>
<td>Systems</td>
<td>Total # Posts</td>
<td>System</td>
<td>NEG</td>
<td>NEU</td>
</tr>
<tr>
<td>---------</td>
<td>--------------</td>
<td>--------</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>18471</td>
<td>Radian6</td>
<td>WB</td>
<td>12.54%</td>
<td>79.68%</td>
</tr>
<tr>
<td>18352</td>
<td>Semantria</td>
<td>WB</td>
<td>30.17%</td>
<td>44.30%</td>
</tr>
<tr>
<td>18352</td>
<td>Semantria</td>
<td>NB</td>
<td>30.17%</td>
<td>46.00%</td>
</tr>
<tr>
<td>18466</td>
<td>Feeling Meter</td>
<td>WB</td>
<td>19.59%</td>
<td>27.99%</td>
</tr>
<tr>
<td>18466</td>
<td>Feeling Meter</td>
<td>NB</td>
<td>18.02%</td>
<td>20.05%</td>
</tr>
<tr>
<td>18469</td>
<td>Sentistrength</td>
<td>WB</td>
<td>37.66%</td>
<td>36.87%</td>
</tr>
<tr>
<td>18469</td>
<td>Sentistrength</td>
<td>NB</td>
<td>38.00%</td>
<td>37.28%</td>
</tr>
<tr>
<td>18420</td>
<td>Stanford</td>
<td>NB</td>
<td>92.23%</td>
<td>3.29%</td>
</tr>
<tr>
<td>36190</td>
<td></td>
<td>WB</td>
<td>62.07%</td>
<td>26.40%</td>
</tr>
</tbody>
</table>

**Sum of percentage points of variation between systems and coders (C123)**

<table>
<thead>
<tr>
<th>Systems</th>
<th>System</th>
<th>Neg-HA</th>
<th>Neg-Sys</th>
<th>Neu-HA</th>
<th>Neu-Sys</th>
<th>Pos-HA</th>
<th>Pos-Sys</th>
<th>Diff-NEG</th>
<th>Diff-NEU</th>
<th>Diff-POS</th>
<th>Diff-POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>18471</td>
<td>Radian6</td>
<td>5252.825175</td>
<td>2316</td>
<td>10098.82318</td>
<td>14717</td>
<td>2890.899101</td>
<td>1382</td>
<td>2936.825175</td>
<td>-4408.176823</td>
<td>1508.899101</td>
<td>127.49%</td>
</tr>
<tr>
<td>18352</td>
<td>Semantria</td>
<td>5218.983683</td>
<td>5562</td>
<td>10242.40826</td>
<td>8130</td>
<td>2872.274392</td>
<td>4660</td>
<td>3343.983683</td>
<td>-5371.274392</td>
<td>2112.408258</td>
<td>23.12%</td>
</tr>
<tr>
<td>18352</td>
<td>Semantria</td>
<td>5218.983683</td>
<td>5536</td>
<td>10242.40826</td>
<td>8258</td>
<td>2872.274392</td>
<td>4958</td>
<td>3343.983683</td>
<td>-5371.274392</td>
<td>3142.408258</td>
<td>127.49%</td>
</tr>
<tr>
<td>18466</td>
<td>Feeling Meter</td>
<td>5251.403263</td>
<td>3629</td>
<td>10306.03263</td>
<td>4036</td>
<td>2890.11655</td>
<td>1084</td>
<td>1622.403263</td>
<td>-1268.61655</td>
<td>8130</td>
<td>2872.274392</td>
</tr>
<tr>
<td>18466</td>
<td>Feeling Meter</td>
<td>5251.403263</td>
<td>3327</td>
<td>10306.03263</td>
<td>3703</td>
<td>2890.11655</td>
<td>11439</td>
<td>1924.403263</td>
<td>-6976.11655</td>
<td>6603.032634</td>
<td>92.47%</td>
</tr>
<tr>
<td>18469</td>
<td>Sentistrength</td>
<td>5252.25641</td>
<td>6955</td>
<td>10307.70696</td>
<td>8609</td>
<td>2890.586081</td>
<td>4705</td>
<td>-1702.74359</td>
<td>3498.70696</td>
<td>7015.88469</td>
<td>37.99%</td>
</tr>
<tr>
<td>18469</td>
<td>Sentistrength</td>
<td>5252.25641</td>
<td>7018</td>
<td>10307.70696</td>
<td>6866</td>
<td>2890.586081</td>
<td>4666</td>
<td>-1765.74359</td>
<td>3422.70696</td>
<td>6863.88469</td>
<td>37.16%</td>
</tr>
<tr>
<td>18420</td>
<td>Stanford</td>
<td>NB</td>
<td>5252.39641</td>
<td>7018</td>
<td>10307.70696</td>
<td>6866</td>
<td>2890.586081</td>
<td>4666</td>
<td>-1765.74359</td>
<td>3422.70696</td>
<td>6863.88469</td>
</tr>
</tbody>
</table>

**C123**

<table>
<thead>
<tr>
<th>Systems</th>
<th>System</th>
<th>Neg-HA</th>
<th>Neg-Sys</th>
<th>Neu-HA</th>
<th>Neu-Sys</th>
<th>Pos-HA</th>
<th>Pos-Sys</th>
<th>Diff-NEG</th>
<th>Diff-NEU</th>
<th>Diff-POS</th>
<th>Diff-POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>18471</td>
<td>Radian6</td>
<td>5252.825175</td>
<td>2316</td>
<td>10098.82318</td>
<td>14717</td>
<td>2890.899101</td>
<td>1382</td>
<td>2936.825175</td>
<td>-4408.176823</td>
<td>1508.899101</td>
<td>127.49%</td>
</tr>
<tr>
<td>18352</td>
<td>Semantria</td>
<td>5218.983683</td>
<td>5562</td>
<td>10242.40826</td>
<td>8130</td>
<td>2872.274392</td>
<td>4660</td>
<td>3343.983683</td>
<td>-5371.274392</td>
<td>2112.408258</td>
<td>23.12%</td>
</tr>
<tr>
<td>18352</td>
<td>Semantria</td>
<td>5218.983683</td>
<td>5536</td>
<td>10242.40826</td>
<td>8258</td>
<td>2872.274392</td>
<td>4958</td>
<td>3343.983683</td>
<td>-5371.274392</td>
<td>3142.408258</td>
<td>127.49%</td>
</tr>
<tr>
<td>18466</td>
<td>Feeling Meter</td>
<td>5251.403263</td>
<td>3629</td>
<td>10306.03263</td>
<td>4036</td>
<td>2890.11655</td>
<td>1084</td>
<td>1622.403263</td>
<td>-1268.61655</td>
<td>8130</td>
<td>2872.274392</td>
</tr>
<tr>
<td>18466</td>
<td>Feeling Meter</td>
<td>5251.403263</td>
<td>3327</td>
<td>10306.03263</td>
<td>3703</td>
<td>2890.11655</td>
<td>11439</td>
<td>1924.403263</td>
<td>-6976.11655</td>
<td>6603.032634</td>
<td>92.47%</td>
</tr>
<tr>
<td>18469</td>
<td>Sentistrength</td>
<td>5252.25641</td>
<td>6955</td>
<td>10307.70696</td>
<td>8609</td>
<td>2890.586081</td>
<td>4705</td>
<td>-1702.74359</td>
<td>3498.70696</td>
<td>7015.88469</td>
<td>37.99%</td>
</tr>
<tr>
<td>18469</td>
<td>Sentistrength</td>
<td>5252.25641</td>
<td>7018</td>
<td>10307.70696</td>
<td>6866</td>
<td>2890.586081</td>
<td>4666</td>
<td>-1765.74359</td>
<td>3422.70696</td>
<td>6863.88469</td>
<td>37.16%</td>
</tr>
<tr>
<td>18420</td>
<td>Stanford</td>
<td>NB</td>
<td>5252.39641</td>
<td>7018</td>
<td>10307.70696</td>
<td>6866</td>
<td>2890.586081</td>
<td>4666</td>
<td>-1765.74359</td>
<td>3422.70696</td>
<td>6863.88469</td>
</tr>
</tbody>
</table>