Comparing the Performance of Knowledge-Based and Machine-learning Approaches for the Detection of Emotions in an English Text

Jeremy Claude Barnes

Màster: Lingüística Teòrica i Aplicada


Directors: Dr. Juan María Garrido Almiñana Dr. Antoni Badia Cardús

Any de defensa: 2015

Col·lecció: Treballs de fi de màster

Programa oficial de postgrau
"Comunicació lingüística i mediació multilingüe"

Departament de Traducció i Ciències del Llenguatge
Comparing the Performance of Knowledge-Based and Machine-learning Approaches for the Detection of Emotions in an English Text

Thesis for the achievement of the degree of MA in Theoretical and Applied Linguistics

Author: Jeremy Barnes

Supervisors: Dr. Juan María Garrido Almiñana

Dr. Antoni Badia i Cardús

July 2015
This page intentionally left blank.
Acknowledgments

My warmest thanks to my tutors Juan María Garrido and Toni Badia for their support and guidance, which have been indispensable. I would like to thank Jessie Walker, who dedicated time and effort to the annotation of the corpus. I would like to thank my wife, Itziar, who has helped and pushed me more than anyone.
# Table of Contents

*Abstract* ............................................................................................................................ 5

1. **Introduction** .................................................................................................................... 5

2. **Goals** ............................................................................................................................. 7

3. **Previous Work** ............................................................................................................... 8
   3.1. **Emotion Theory** ........................................................................................................ 8
   3.2. **Knowledge-Based Approaches** ................................................................................. 9
   3.3. **Machine-Learning Approaches** ............................................................................. 11
       3.3.1. Naive Bayes ........................................................................................................... 12
       3.3.2. Maximum Entropy ................................................................................................ 12
       3.3.3. Support Vector Machine ....................................................................................... 12
       3.3.4. Implementations ................................................................................................... 13

4. **Corpus and Markup** ....................................................................................................... 13
   4.1. **NPS Corpus** ............................................................................................................. 13
   4.2. **Annotation of Corpus** ............................................................................................ 14
   4.3. **Preprocessing** ......................................................................................................... 19

5. **Experimental Design** .................................................................................................... 20
   5.1. **Knowledge-Based Approach** .................................................................................. 20
       5.1.1. Extraction of Sentiment Lexicon ........................................................................... 20
       5.1.2. EmotionFinder English ......................................................................................... 21
   5.2. **Machine-Learning Approach** .................................................................................. 23
       5.2.1. Naive Bayes Classifier ........................................................................................ 23
       5.2.2. Maximum Entropy Classifier .............................................................................. 23
       5.2.3. Support Vector Machine Classifier ..................................................................... 23

6. **Results** ......................................................................................................................... 23
ABSTRACT

The detection of emotion and sentiment analysis are very hot topics at the moment and the detection of emotion from written text still remains a difficult subject of this area of research. The main approaches to this task are knowledge-based approaches and machine-learning approaches. This paper examines the performance of two approaches (a knowledge-based and a machine-learning approach) on a small corpus of chat text annotated with emotion labels. It will be shown that the machine-learning approach used in this experiment outperforms the knowledge-based approach in all aspects.

Keywords

Emotion detection, sentiment analysis, machine-learning, knowledge-based, corpus annotation

1. INTRODUCTION

Detecting emotions in text is a practice that people do every day to a high level of precision, normally without much effort. Whether it be reading Tripadvisor for hotel reviews or an opinion piece in a newspaper, humans are constantly looking for emotion-related information. The amount of opinionated data available, however, means that it is quite easy to become overwhelmed when looking for information. For this reason, the idea of computationally analyzing emotion is very interesting.

Sentiment analysis and emotion detection are two lines of research that are related, yet have slightly different goals. Sentiment analysis tends to be more interested in discovering a small number of categories (positive vs. negative or 1-5 star-ratings) and does this from
very coarse (document level) to very fine (feature level). It tends to be associated to NLP or data mining groups, as these are the ones that do the most research. Emotion detection, however, is usually more interested in finding a larger number of emotions (discrete emotions, or levels of valence and arousal) for the purpose of Affective Computing (Picard 1997) or Text-To-Speech (Garrido et al. 2012). Both of these areas of research, nonetheless, overlap in many ways and it is beneficial that they share information, as they are complementary in many ways.

The two main approaches to sentiment analysis and emotion detection are knowledge-based or machine-learning approaches. Although knowledge-based approaches began long earlier, machine-learning is the predominant way to conduct sentiment analysis. Knowledge-based systems are known for being difficult and expensive to build, as they require more human input (in the forms of sentiment lexicons and rules) in order to perform well. Machine-learning techniques used are usually a case of supervised learning, in which a classifier must be trained with previously annotated data. This data also requires more time, effort and money than any other part of the research, which has led some researchers to try semi-supervised and unsupervised techniques (Turney 2002). Machine-learning has also been shown to be heavily domain dependent, which is less of a problem for knowledge-based approaches.

The analysis of informal text (such as product reviews, blogs, chat text or WhatsApp messages) is increasingly important, as the number of Internet users that post opinionated and emotion-related data has grown since the beginning of Web 2.0. This kind of text presents several problems not found in traditional text, such as the common use of abbreviations, emoticons, unorthodox spelling and grammar and a high level of creativity
with language. All of these provide an extra challenge for sentiment analysis and emotion detection.

The contributions of this paper are as follows: First, related work is discussed in detail. Second, it outlines the annotation of a chat corpus with a large number of emotion tags. Next, it introduces a knowledge-based system and machine-learning system in detail for the classification of emotions. Then, the performance of both systems is compared on various levels, from their performance on all 33 emotions with strengths to the most basic positive, negative, neutral tags in order to investigate which system works best at which level of granularity. Finally, it will be shown that the machine-learning approach proves more accurate than the knowledge-based system on this corpus.

2. GOALS

One of the main drawbacks of current knowledge-based and machine-learning approaches is the difficulty of annotating large amounts of data in short amount of time. The amount of time and effort required to annotate a large corpus and then analyze the data is often prohibitive. For this reason, discovering which current models are better adapted to working with small data sets is one of the major motivations for this research
3. PREVIOUS WORK

3.1 What is Emotion

The term emotion is often confusing, as it is often used interchangeably with mood, state of arousal or attitude. Cowie & Cornelius (2003) point out that “In the meaning that people tend to feel is central, emotions are episodes that are relatively brief and highly distinctive. There is no generally agreed, compact term for these episodes.” It is therefore difficult to give an exact definition of what an emotion is. In fact, there are many theoretical models of emotion, none of which agree on what a basic set of emotions would be.

Research into sentiment analysis and emotion detection has often lacked a theoretical basis for what constitutes an emotion. This stems mainly from the fact that researchers in these areas were more interested in creating practical applications, rather than theoretically sound ones (Cowie et al 2001). Calvo & D'Mello (2010) argue, however, that it is essential to incorporate emotion theory into the designs of Affective Computing systems, as cross-disciplinary collaboration will be necessary to improve the efficacy of these.

Research on emotion detection, especially sentiment analysis, has also tended to choose a small number of emotion labels. Several researchers have suggested, nevertheless, that such a narrow selection of emotions may not be enough for affective computing and emotion detection (Calvo & D'Mello 2010, Garrido et al. 2012, Paltoglou 2013). In line with the latter, the research in this paper attempts to explore a larger number of emotion labels in order to judge if it is feasible to do so.
3.2 Knowledge-Based Approaches

Knowledge-based approaches were the first to be used for sentiment analysis and emotion detection. The most straightforward are those that use the appearance or counts of keywords to determine the emotion of a text. The largest amount of work in this approach is the collection of these keywords, which are then used to build a keyword-dictionary or lexicon. There have been many attempts at creating such a keyword-dictionary for the purpose of emotion detection, from the General Inquirer (Stone et al. 1966), which was among the earliest attempts to assemble a sentiment lexicon, to the many affective lexicons based on Wordnet, such as Wordnet-Affect (Strapparava & Valitutti 2004), SentiWordnet (Esuli & Sebastiani 2006) or Q-Wordnet (Agerri & Garcia-Serrano 2010).

Yet a simple keyword-based system has many pitfalls that are well-known. Negation often has the effect of completely changing the emotion of a phrase, as in the following example:

Ex 1.

I'm happy → joy
I'm not happy → sadness

Das & Chen (2001) incorporated heuristic methods to deal with negation. They used a simple technique in which in sentences where a negation element was present ('not', 'never', 'no), the words following the negation were tagged with a negation marker ('--n'). This allowed them to correctly label sentences such as example one.
Unfortunately, it is not always the case that the occurrence of a negation word leads to the exact opposite emotion, as the following example shows:

**Ex 2.**

I wonder if John isn't happy → interest/wonder

Besides negation, enhancers such as ‘very’ or ‘extremely’ are often used to determine a more fine-grained level of emotion, boosting an emotion tag from a lower strength to a higher, or vice-versa.

**Ex. 3**

I’m really worried about my exam → strong worry

I’m a little worried about my exam → weak worry

A very different approach which Balahur et al. (2012) took was to construct a knowledge-based system based on ontologies, which they called EmotiNet. They first created an ontology core and then populated it semi-automatically with examples from a corpus. When compared to the GENESIS system (Scherer 1993), the technique used in EmotiNet proved to be quite robust to domain changes, but the precision and recall measures remained quite low.

Strapparava & Mihalcea (2008) constructed a corpus of newspaper headlines which was annotated with 6 basic emotions and also degree of emotional load, ranging from 0 (a lack of emotional load) to 100 (maximal emotional load). They then used Latent Semantic Analysis, a technique by which text is represented as a vector space model, and WordNet-Affect to classify the headlines. They also used a Naive Bayes classifier trained on blog
posts to compare the results. The conclusion was that the system based on WordNet-Affect had the most precision, at the cost of low recall. Compared to the Naive Bayes classifier, the LSA model proved to have significantly better recall and F1 scores for most emotion categories.

Kolz et al. (2014) introduced EmotionFinder, which is system that uses an emotional dictionary and a set of rules to assign one of nine emotional labels (including 'Neutral') and a strength (from 1 to 3) to sentences for the purpose of enhancing a processing module for Text-to-Speech, TexAFon (Garrido et al. 2012). EmotionFinder showed some improvement over similar systems with fewer emotion categories. However, it was also clear from this paper that the system was highly dependent on the corpus used in its creation.

### 3.3 Machine-learning approaches

The machine-learning approaches to sentiment analysis have proliferated in the last decade, leading to the publishing of thousands of articles ranging from different algorithms to feature selection or techniques for dimensionality reduction. For a complete survey of machine learning techniques related to sentiment analysis, see Pang & Lee (2006).

Some of the most popular machine learning algorithms that are used in sentiment analysis will be discussed in the following three sections.
3.3.1 Naive Bayes Classifier

A Naive Bayes classifier is a method of supervised learning, based on Bayes’ theorem. The name ‘naive’ is derived from that fact that it makes unrealistic assumptions about the independence of variables. Nonetheless, it has proven to be a surprisingly accurate classifier in many experiments (Pang et al. 2002). To formalize this classifier at sentence level, given a set of features X representing a sentence and a set of emotion labels L:

\[ \text{Emotion Label} = \arg \max P(\text{label} \mid X) \]

Figure 1.

3.3.2 Maximum Entropy Classifier

A Maximum Entropy classifier is a classifier that considers all of the probability distributions that are empirically consistent with the training data and chooses the distribution with the highest entropy (Bird et al. 2009). First introduced to NLP by Berger et al. (1996), it has been used successfully in a number of implementations (Pang et al. 2002, Manning & Klein 2003). One of the advantages of this algorithm is that it allows for the incorporation of a large amount of linguistic data.

3.3.3 Support Vector Machine

Support Vector Machines (SVM) are state-of-the-art in many areas of NLP (Pang et al. 2002). They are non-probabilistic discriminative models which create a hyperplane between examples to separate classes.
3.3.4 Implementations

In Pang & Lee (2002), the authors compared the performance of Naive Bayes, Maximum entropy and Support Vector Machine classifiers on movie review corpus. The reviews were labeled as positive and negative and they used. After comparing all of the algorithms trained on various features, they found that the SVM trained on unigrams performed the best, at 82.9 percent accuracy.

C. Alm, D. Roth & R. Sproat (2005) proposed a machine-learning system for the identification of neutral, positive and negative valence, which they cast as a multiple-class classification problem. Their motivation for using a larger range of emotional categories was to render Text-to-Speech systems more lifelike. Their approach resulted in an F score of .69 in the neutral category, .39 in the negative and .13 in the positive. It showed however that even a small corpus could be used for machine-learning purposes.

4. CORPUS AND MARKUP

4.1 NPS Chat Corpus

The corpus chosen for this experiment was the NPS chat corpus, collected by Jane Lin (Lin 2007) and further processed by Eric N. Forsythe and Craig H. Martell (2007) This corpus was taken from chat posts from a number of different chat rooms. Lin's goal was to analyze the chat posts in order to automatically profile authors of the posts.

Forsythe continued processing the corpus through POS tagging, parsing and discourse information which they did both automatically and by hand.
For this corpus, a subset of the original corpus was used. 2000 posts were taken from two different age-oriented chat rooms, one (10-19-20s 706) for young adults and the other (10-26-teens 706) for teenagers. As mentioned in Section 2, the small amount of annotated data was due to a lack of time and resources. Yet, since this is one of the motivating factors of this research, we believe that this can be of interest.

4.2 Annotation of the Corpus

Two annotators were chosen to annotate the corpus separately. Both annotators were native speakers of American English between 30 and 32 years old and both had extensive experience in the chat domain. The annotators were given no explicit training and were told to follow their instincts in order to complete the annotations.

For the annotation of this corpus, following Garrido et al. (2014), we have decided to use a set of emotion labels based on those put forth by the HUMAINE project’s EARL proposal. The Human-Machine Interaction Network on Emotion (HUMAINE) project is a network whose aim is to “lay the foundations for European development of systems that can register, model and/or influence human emotional and emotion-related states and processes.” (Douglas-Cowie, Cox et al., HUMAINE deliverable D5f) Finally, a subset of 33 emotions in the EARL proposal were taken as being more relevant to the chat domain (in this experiment the 9 physiological states which were included in Garrido’s work were not considered, as an initial test proved that there were very few examples of these in this particular corpus.) Following Garrido et al. (2012) and Kolz et al. (2014) we also included strength labels, from 1-3, depending on the intensity of the emotion. The proposed emotion labels are found in Table 1.
Each post from the chat corpus was annotated with an emotion from the list of 33 emotions in table 1 and a strength from 1 to 3, 1 being a weak emotion and 3 being a strong emotion.

A third annotator was chosen to disambiguate in those instances that the first two annotators did not agree. The inter annotator agreement between all three annotators is shown below in table 2. The composition of the corpus is shown below.

The results of the annotation process are discussed below. Table 2 shows the observed agreement between all three annotators, while table 3 gives the Cohen's Kappa coefficients of the first two annotators.
The inter annotator agreement may seem quite low, yet there are many reasons for the low amount of agreement. For many of the examples found in the corpus, several emotion tags could have easily been applied correctly, such as in the following example.

**Ex. 4**

User33: lol :P

Here, <fun>, <joy> and <mockery> could all easily be considered correct tags. Yet, because of the decision to give a single tag to each sentence in this experiment, the annotators could only choose one. This, therefore, leads to a lower degree of agreement.
than if annotators had been allowed to give several tags and agreement had been based on coinciding on at least one tag (Neviarouskaya et al. 2007).

The corpus is not balanced in the sense that there are many more examples of the most common labels than the least common. The numbers of each emotion label are given in table 4 below, from most common to least common.

<table>
<thead>
<tr>
<th>Label</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>interest</td>
<td>520</td>
</tr>
<tr>
<td>fun</td>
<td>417</td>
</tr>
<tr>
<td>disapproval</td>
<td>100</td>
</tr>
<tr>
<td>affection, complicity, mockery</td>
<td>99 - 50</td>
</tr>
<tr>
<td>anger, excitement, indifference, pride</td>
<td>50 - 30</td>
</tr>
<tr>
<td>irony, doubt, disappointment, worry, trust</td>
<td>29-20</td>
</tr>
<tr>
<td>boredom, distrust, resignation, shame, guilt, impatience, compassion, admiration</td>
<td>19-10</td>
</tr>
<tr>
<td>dejection, envy, nostalgia, sadness, disgust, impotence, relief</td>
<td>&lt;10</td>
</tr>
</tbody>
</table>

Table 4. Occurrences of Emotion Labels

Besides the complete group of emotions seen in Table 1, we also used a set of compacted emotions, based on the analysis of many mistakes in the annotation process. The most commonly confused emotions were joy and fun (171 discrepancies) and interest and complicity (42 discrepancies). After analyzing the data, we decided to combine these two
groups of emotions into one label each. This reduced emotion labels are found in bold in Table 5.

<table>
<thead>
<tr>
<th>Positive</th>
<th>Neutral</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Affection/Complicity</strong></td>
<td>Surprise</td>
<td>Mockery</td>
</tr>
<tr>
<td><strong>Joy/Fun</strong></td>
<td>None</td>
<td>Irony</td>
</tr>
<tr>
<td>Excitement</td>
<td></td>
<td>Boredom</td>
</tr>
<tr>
<td>Interest</td>
<td></td>
<td>Doubt</td>
</tr>
<tr>
<td>Pride</td>
<td></td>
<td>Distrust</td>
</tr>
<tr>
<td>Relief</td>
<td></td>
<td>Dejection</td>
</tr>
<tr>
<td>Compassion</td>
<td></td>
<td>Disappointment</td>
</tr>
<tr>
<td>Admiration</td>
<td></td>
<td>Resignation</td>
</tr>
<tr>
<td>Trust</td>
<td></td>
<td>Worry</td>
</tr>
<tr>
<td>Surprised</td>
<td></td>
<td>Envy</td>
</tr>
<tr>
<td>Mockery</td>
<td></td>
<td>Disapproval</td>
</tr>
<tr>
<td>None</td>
<td></td>
<td>Nostalgia</td>
</tr>
<tr>
<td>Pride</td>
<td></td>
<td>Shame</td>
</tr>
<tr>
<td>Surprised</td>
<td></td>
<td>Guilt</td>
</tr>
<tr>
<td>None</td>
<td></td>
<td>Sadness</td>
</tr>
<tr>
<td>Pride</td>
<td></td>
<td>Disgust</td>
</tr>
<tr>
<td>Surprised</td>
<td></td>
<td>Impatience</td>
</tr>
<tr>
<td>None</td>
<td></td>
<td>Anger</td>
</tr>
<tr>
<td>Pride</td>
<td></td>
<td>Impotence</td>
</tr>
<tr>
<td>Surprised</td>
<td></td>
<td>Fear</td>
</tr>
</tbody>
</table>

Table 5. A list of the compacted emotions.

Next, following Kolz et al. (2014) we removed those emotions that did not amount to at least 10 percent of the total number of labels, leaving the emotion labels in Table 6.

<table>
<thead>
<tr>
<th>Positive</th>
<th>Neutral</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Affection/Complicity</strong></td>
<td>Surprise</td>
<td>Mockery</td>
</tr>
<tr>
<td><strong>Joy/Fun</strong></td>
<td>None</td>
<td>Irony</td>
</tr>
<tr>
<td>Excitement</td>
<td></td>
<td>Boredom</td>
</tr>
<tr>
<td>Interest</td>
<td></td>
<td>Doubt</td>
</tr>
<tr>
<td>Pride</td>
<td></td>
<td>Distrust</td>
</tr>
<tr>
<td>Relief</td>
<td></td>
<td>Dejection</td>
</tr>
<tr>
<td>Compassion</td>
<td></td>
<td>Disappointment</td>
</tr>
<tr>
<td>Admiration</td>
<td></td>
<td>Resignation</td>
</tr>
<tr>
<td>Trust</td>
<td></td>
<td>Worry</td>
</tr>
<tr>
<td>Surprised</td>
<td></td>
<td>Envy</td>
</tr>
<tr>
<td>Mockery</td>
<td></td>
<td>Disapproval</td>
</tr>
<tr>
<td>None</td>
<td></td>
<td>Nostalgia</td>
</tr>
<tr>
<td>Pride</td>
<td></td>
<td>Shame</td>
</tr>
<tr>
<td>Surprised</td>
<td></td>
<td>Guilt</td>
</tr>
<tr>
<td>None</td>
<td></td>
<td>Sadness</td>
</tr>
<tr>
<td>Pride</td>
<td></td>
<td>Disgust</td>
</tr>
<tr>
<td>Surprised</td>
<td></td>
<td>Impatience</td>
</tr>
<tr>
<td>None</td>
<td></td>
<td>Anger</td>
</tr>
</tbody>
</table>

Table 6. A list of the compacted emotions.
Finally, we also used the simple positive, negative and neutral schema. The specific emotion labels were mapped to each of their more general categories. We will compare the effectiveness of the different approaches on all five sets of labels.

Other researchers have also commented on the difficulty of inter annotator agreement for the task of emotion detection (Neviarouskaya et al. 2007, Scherer 1982). While performing research with the Leeds-Reading Database Greasely et al. (2000) commented on the difficulty of recognizing emotion for human annotators. In their experiment, subjects listened to speech samples drawn from this database and applied one of five labels (happiness, sadness, anger, fear, disgust). Even with this reduced set of labels, the annotators agreed on only 54 percent of the negative labels. In our corpus, with the reduced list of fifteen emotion labels, the inter annotation agreement was similar. This leads us to the conclusion that at least at the level of the reduced list of emotions, the agreement level is acceptable.

4.3 PREPROCESSING

Thelwall et al. (2010) commented on the importance and difficulty of preprocessing texts from the chat domain, pointing out how important it is to incorporate ways to handle abbreviations, emoticons and non-standard spelling in order to gain accuracy. During the preprocessing of this corpus, all common contractions were expanded. It was then tokenized using regular expressions. Several important tokens that are common in the chat domain, such as emoticons and questions and answers about someone’s age, sex and location (a/s/l), were kept as one token, as they provide important information about the interlocutor’s emotional state.
The corpus was then POS-tagged using the backoff Trigram tagger available from NLTK, which was trained on the Brown corpus. The POS-tagging provided a rudimentary word-sense disambiguation, especially for nouns, verbs and adjective.

The POS-tagged version was then used to lemmatize the corpus with the intention of being able to generalize much better, since Kolz et al (2014) concluded that lemmatizing the corpus led to significantly better results using their knowledge-based approach. This was done using NLTK’s WordNet Lemmatizer.

5. EXPERIMENTAL DESIGN

5.1 Knowledge-Based Approach

5.1.1 Extraction of Sentiment Lexicon

Because of the decision to use a large number of emotion labels, the decision was made to extract a sentiment lexicon from the labeled corpus rather than attempt to modify an existing sentiment lexicon. As mentioned above, the corpus was POS-tagged using a backoff trigram tagger available in NLTK. It was then lemmatized using NLTK’s Wordnet Lemmatizer. Unigrams and bigrams that were associated with each emotion label were extracted along with the number of times they co-occurred with each label. Stopwords were removed. The number of co-occurrences were then used as a cutoff (the minimum was set to 3 occurrences) to build the sentiment lexicon.

The final sentiment lexicon consisted of a list of terms with the emotion label, the strength associated with it, and the number of occurrences.
5.1.2 EmotionFinder English

A program called EmotionFinder English was used to test the knowledge-based approach in this experiment. It is based on EmotionFinder, which is a program that was developed by Kolz et al. (2014) to detect emotions for TexAFon. It incorporates a dictionary of lemmas and weights, which are the main features taken into consideration when deciding on the emotion of a sentence. This dictionary is automatically extracted from the gold standard corpus. It also includes a series of rules, which include the detection of negation elements, booster words, emoticons and information related to age/sex/location, such as “20/m/Texas.”

The negation modules uses regular expressions to look for occurrences of ‘no’ or ‘not’ in a sentence. Take the sentence “I am not particularly interested.” If the original label given to the utterance were “<Interest>” based simply on the the word “interested” in the sentence, the negation module would then find the negation and assign the label “<Indifference>”. Of course, not all emotions have an exact opposite. In these cases the emotion label was set to “<None>” as a default.

The booster word module works on the same concept and uses the words “very”, “really” and “a lot” to increase the strength label of an emotion by one point.
The emoticon module uses regular expressions to search for typical emoticons, such as the smiley face “:-)” or sad face “:-(“. If one is found, this is taken to be a significant sign of the overall sentiment and the emotion label is set directly to the emotion associated with the emoticon.

Finally, the a/s/l module looks for any questions or answers about someone’s age, sex and location. This was taken as an instance of “<Interest>”, since it was almost always associated with at least a small interest in other chatters.

Figure 2. Work Flow of EmotionFinder English
5.2 Machine-learning Approach

5.2.1 Naive Bayes Classifier

For this experiment, the Bernoulli Naive Bayes classifier from Scikit-learn (Pedregosa et al. 2011) was implemented in Python using the ‘bag-of-words’ approach. The features used to train the classifier were unigrams, bigrams, trigrams, the presence of emoticons and the presence of age/sex/location related information (20/m/texas).

5.2.2 Maximum Entropy Classifier

For this experiment, the Maxent classifier from NLTK (Bird et al. 2009) was trained on the same features as the Naive Bayes classifier mentioned in the previous section using generalized iterative scaling (IIS) and a cutoff of .1 in the minimum log likelihood difference between iterations.

5.2.4 Support Vector Machine

For this experiment, the SMO algorithm from Scikit-learn with a linear kernel was trained on the same features as the other classifiers.

6. RESULTS

Each algorithm's accuracy was tested with a ten fold cross-validation, in which the corpus was divided into ten even parts and the algorithm was trained on nine sections and tested on the tenth. This is a technique which allows one to test whether a classifier is overfitting, a problem which occurs frequently with small data sets (Kohavi 1995). The final accuracies and F measures given below are the average of the ten accuracies and F measures.
<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>All emotions and strength</td>
<td>27.95%</td>
<td>.32</td>
<td>.05</td>
<td>.08</td>
</tr>
<tr>
<td>All emotions without strength</td>
<td>35.75%</td>
<td>.53</td>
<td>.06</td>
<td>.11</td>
</tr>
<tr>
<td>Compacted emotions</td>
<td>37.2%</td>
<td>.52</td>
<td>.07</td>
<td>.11</td>
</tr>
<tr>
<td>Threshold emotions</td>
<td>37.05%</td>
<td>.53</td>
<td>.10</td>
<td>.12</td>
</tr>
<tr>
<td>Positive Negative and Neutral</td>
<td>61.90%</td>
<td>.68</td>
<td>.34</td>
<td>.45</td>
</tr>
</tbody>
</table>

**Table 7: Results of Naive Bayes Algorithm**

Variance of f-measures with all emotions and strengths: 0.05-0.1

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>All emotions and strength</td>
<td>34.15%</td>
<td>.29</td>
<td>.09</td>
<td>.14</td>
</tr>
<tr>
<td>All emotions without strength</td>
<td>47.15%</td>
<td>.35</td>
<td>.14</td>
<td>.2</td>
</tr>
<tr>
<td>Compacted emotions</td>
<td>48.2%</td>
<td>.32</td>
<td>.13</td>
<td>.18</td>
</tr>
<tr>
<td>Threshold emotions</td>
<td>55%</td>
<td>.42</td>
<td>.25</td>
<td>.32</td>
</tr>
<tr>
<td>Positive Negative and Neutral</td>
<td>66.45%</td>
<td>.62</td>
<td>.44</td>
<td>.52</td>
</tr>
</tbody>
</table>

**Table 8: Results of Maximum Entropy Algorithm**

Variance of f-measures with all emotions and strengths: .11 - .18

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>All emotions and strength</td>
<td>33.65%</td>
<td>.26</td>
<td>.1</td>
<td>.14</td>
</tr>
<tr>
<td>All emotions without strength</td>
<td>45.6%</td>
<td>.37</td>
<td>.17</td>
<td>.23</td>
</tr>
<tr>
<td>Compacted emotions</td>
<td>47.5%</td>
<td>.36</td>
<td>.16</td>
<td>.22</td>
</tr>
<tr>
<td>Threshold emotions</td>
<td>54.65%</td>
<td>.46</td>
<td>.29</td>
<td>.36</td>
</tr>
<tr>
<td>Positive Negative and Neutral</td>
<td>64.85%</td>
<td>.53</td>
<td>.5</td>
<td>.52</td>
</tr>
</tbody>
</table>

**Table 9: Accuracy of SVM Algorithm**

Variance of f-measures with all emotions and strengths: .15 - .3
<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>All emotions and strength</td>
<td>21%</td>
<td>.1</td>
<td>.06</td>
<td>.08</td>
</tr>
<tr>
<td>All emotions without strength</td>
<td>21.5%</td>
<td>.12</td>
<td>.06</td>
<td>.08</td>
</tr>
<tr>
<td>Compacted emotions</td>
<td>36.5%</td>
<td>.17</td>
<td>.09</td>
<td>.12</td>
</tr>
<tr>
<td>Threshold emotions</td>
<td>37.6%</td>
<td>.17</td>
<td>.09</td>
<td>.12</td>
</tr>
<tr>
<td>Positive Negative and Neutral</td>
<td>30.5%</td>
<td>.13</td>
<td>.8</td>
<td>.1</td>
</tr>
</tbody>
</table>

Table 10: Accuracy of EmotionFinder English

As we can see from the data in this section, the machine learning techniques outperform the knowledge-based techniques on all of the emotion labels. This is especially noticeable at the coarser levels of positive, negative and neutral. The reason for this is that EmotionFinder English was originally designed with a lexicon and rules to create a finer grained analysis. The use of the same algorithms with a small wrapper to translate the output label is clearly not a substitute for specific rules and lexicon at each level of granularity. Machine-learning algorithms may also be able to pick up more abstract patterns of cooccurrence that are useful at coarser levels.

The Maxent classifier had better accuracy than the other two, but its precision, recall and F measures were lower. The Naive Bayes classifier, had very high precision scores, but at the expense of recall. The SVM classifier is the most balanced, with recall and F measures that outperform the others at most levels. This seems to confirm the usefulness of SVMs in the classification of emotions.
7. CONCLUSION AND DISCUSSION

In this paper we have compared the performance of machine-learning and knowledge-based algorithms on a small chat corpus, having proven that the machine-learning techniques outperformed the knowledge-based technique. We have also compared their level of performance when given different emotional labels.

From the results of both approaches it is clear that the complete set of labels with strengths is too ambitious at this moment. Even the relatively good results in accuracy (34.15% with the MaxEnt classifier) are more likely due to overfitting than to a real ability to discriminate these labels.

For the future, comparing the performance of these algorithms on a different domain would be another useful experiment, as evidence suggests that machine-learning algorithms are very much limited to their domain. Creating a more domain-independent sentiment lexicon for EmotionFinder English would also be beneficial, as it would improve its performance.

On another note, an improved method of normalization and tokenization for this domain is critical, as a large amount of information was lost due to spelling errors and the creativity with language that one finds in chat rooms.
8. References


E. Douglas-Cowie, Cox et al. HUMAINE D5f deliverable; http://emotion-research.net/download/pilot-db/ [last access: 7.7.2015].


