Thesis to obtain a Master degree at the Shanghai Jiaotong University

Automatic Humor Classification on Twitter

Author	Yishay Raz
Supervisor	Professor Li Fang
Research direction	Text Classification
Thesis date	December 2013

Dedication

To my beloved parents, Dalia and Yossi, who stood by me, and tried their best to help in the months of writing this thesis.

To Danor, who made this two-year journey enjoyable.

Acknowledgements

Thanks and appreciation go to Prof. Li Fang who encouraged me to choose a topic close to me for this thesis, and helped me with advice and reference papers.

Thanks to the LCT acceptance committee that gave me this valuable opportunity and interesting international experience.

Dr. Valia Kordoni was also very valuable and taught me so much about myself and the world we live in.

Finally, thanks and love go to my loving parents who gave me the best atmosphere, place and good food, so I could concentrate on writing this thesis.

Abstract

Automatic Humor Classification on Twitter

Much has been written about humor and even sarcasm automatic recognition on Twitter. Nevertheless, the task of classifying humorous tweets according to the type of humor has not been confronted so far, as far as we know.

This research is aimed at applying classification algorithms and other NLP algorithms to the challenging task of automatically identifying the type of humor appearing in messages on Twitter.

The different methods, algorithms, tools and classifiers used are discussed, as well as the specific difficulty encountered due to the very subjective nature of humor and the informal language applied in tweets.

It is shown that the discussed methods improve the accuracy of classification by up to 5% above the baseline which is ZeroR, the algorithm that classifies all instances to the majority class.

Table of Contents

List of Tables

List of Figures

1 Preliminaries

1.1 MOTIVATION

The interaction between humans and machines has long extended out of the usability aspect. Nowadays, computers are not merely a tool to extend our lacking memory and manpower, but also play a stronger role in communication, entertainment and motivation. These aspects may be found in such systems as Chatter-bots, gaming and decision making systems.

Humor is extremely important in any communicative form. It affects not only feelings but also influences human beliefs. It has even been shown to encourage creativity.

Enabling a machine to classify humor types (and in the future also topics) can have many practical applications such as automatic humor subscriptions that send us only those messages that will most probably make us laugh. It can serve as a basis for further research on humor generation of witty and adequate responses by conversational agent applications.

We tend to expose more about ourselves in humor than in regular prose.

In the next section we will highlight several research results from the fields of psychology and sociology that support this argument, and explore the differences in humor produced by different groups. This knowledge can be used to identify the latent attributes of the tweeters, e.g. gender, geographical location or origin and personality features, all based on their tweets. Aggressiveness in humor can be viewed as a potential warning sign and teach us about the author’s mental well-being and malicious intentions.

1.2 OUTLINE

The remainder of the paper is organized as follows: **related work** is reviewed in the next section. Section **Method** briefly describes the model used in the experiments and its implementaion. Section **Experiments** describes the data, tasks and algorithms of humor classification and its results. Section **Conclusion** signs this work and gives ideas for further research.

2 Related Work

We will survey the research work related to our thesis in four different points of reference.

2.1 HUMOR RESEARCH

While the classification of other types of data, and identifying whether tweets are humorous, sarcastic, or neither, has been examined closely in recent years, I am unaware of any research that has been done on automatic humor classification by type or topic.

2.1.1 First Studies – Humor Generation

One of the first studies on computational humor was done by Binsted and Ritchie (1997), in which the authors modeled puns based on semantics and syntax.

This work paved the way for humor generation research works, such as LIBJOG (Raskin and Attardo 1994), JAPE (Binsted and Ritchie 1994, 1997) and HAHAcronym (Stock and Strapparava, 2003). The two former systems were criticized as pseudo-generative because of the template nature of their synthesized jokes. The latter is also very limited in its syntax.

2.1.2 Later Studies – Humor Recognition

Only in later studies was the recognition of humor examined. (Mihalcea and Strapparava 2005) used content and stylistic features to automatically recognize humor.

Very good results were recorded with up to 97% accuracy for one-liners.

This was done, however, on a more homogenous set of data, one-liners, that, unlike tweets, are formal, grammatically correct and often exhibit stylistic features, such as alliterations and antonyms, which were indeed chosen and features but seldom appear in tweets.

Davidov et al. (2010) recognized sarcastic sentences in Twitter. They used a semi-supervised algorithm to acquire features that could then be used by the classifier to decide which data item was sarcastic. In addition to these lexical patterns, the classifier also used punctuation-based features (e.g. number of appearances of "!"). This procedure achieved an F-score of 0.83 on the Twitter dataset.

2.1.3 Humor Theories

There are basically three theories of humor mentioned in related works:

- The incongruity theory
- The superiority theory
- The relief theory.

2.1.3.1 The incongruity theory

This theory suggests that the existence of two contradictory interpretations to the same statement is a necessary condition for humor. It was used as a basis for the Semantic Script-based Theory of Humour (SSTH) (Raskin 1985), and later on the General Theory of Verbal Humour (GTVH) (Attardo and Raskin 1991).

(Taylor 2010) found that the semantic recognition of humor is based on this theory and on humor data that supports it.

We encountered many funny tweets that seem not to comply with this theory.

For example: **There is a new IPod app that translates Jay Leno into funny.**

It appears that some humorous statements can lack any incongruity.

2.1.3.2 The superiority theory

The superiority theory claims that humor is triggered by feelings of superiority with respect to others or to the listener from a prior event (Hobbes 1840).

Our impression is that any humorous text indeed either provokes superiority with respect to the subject of the joke, who is being tarnished, or with respect to the listener himself who feels superior to his position before having been exposed to the joke, because of understanding the witty punch line.

2.1.3.3 The relief theory

The relief theory views humor as a way out of taboo and a license for banned thoughts. Through humor the energy inhibited by social etiquette can be released and bring relief to both the author and audience.

Freud, as early as 1905, supported this theory and connected humor to the unconscious (Freud, 1960).

Minsky (1980) embraces the theory and observes the faulty logic in humor as another steam-releasing trait.

2.1.4 Hurley et al. 2011

(Hurley et al. 2011) try and explain humor differently, as a physiological-evolutional trait that comes to reward our brain in its endeavor to correct its perception errors and misbelieves. Just as the joy of eating and sex see that we do what it takes to survive and procreate, mirth, the joy of humor makes sure that we get a good feeling when we purify our brain and allow it to reassess its errors, and prefer this daily activity upon other less rewarding thoughts.

2.1.5 Mihalcea 2006

(Mihalcea 2006) enumerated the most discriminative content-based features learned by her humor classifier. The more substantial features were found to be human-centric vocabulary, professional communities and human weaknesses that often appear in humorous data. These seem to go hand in hand with the superiority theory. We will show that these discriminative features of humor, more than the three theories mentioned above, will be of greatest value to our task.

2.1.6 Research about humor types and differences

We will now explore what research has been performed on the actual content and types of humor, aside from the computer point of view:

(Hay 1995,2000) describes in her work the difference between humor produced by the two genders.

In the Gender and Humor chapter of her thesis, Hay (1995) surveys old research that claimed women are less inclined towards humor than men. Freud (1905) claimed women do not need a sense of humor because they have fewer strong taboo feelings to repress. This perception is slowly changing, with more contemporaneous work claiming that humor is different between genders. (Hay 1995) concludes that:

- men are more likely to use vulgarity and quotes than women
- women are more likely to use observational humor

To a lesser degree:

- men tend to use more role play and wordplay
- women are more likely to use jocular insults

(Hay 2000) investigates the functions humor serves:

- "Women are much more likely to share funny personal stories to create solidarity."
- "Whereas the men used other strategies to achieve the same goal. They were more likely to reminisce about shared experiences or highlight similarities to create solidarity within the group (of friends)."
- While teasing was used in single-sex groups both to create power and solidarity, this behavior reduced markedly in mixed groups.

(States et al. 1994) describe the main characteristics of Jewish humor, as one that copes with harsh conditions, cherishes Jewish uniqueness, is self-disparaging and expresses Jewish stereotypes. All of these emanate from the unique situation of Jews in the Diaspora.

(Nevo 1984) checked in her study the appreciation and production of humor as an expression of aggression among Jews and Arabs in Israel. Two sets of hypotheses were compared. The ones derived from the social approach claiming that social status determines the expression of aggression in humor and hence members of the Arab minority will express less of it, despite being more frustrated, were the ones excepted.

There are many taxonomies of humor as reviewed by (Hay, 1995), and the one which best suits our data contains the following categories:

1 Anecdotes

2 Fantasy

3 Insult

4 Irony

5 Jokes

6 Observational

7 Quote

8 Role play

9 Self deprecation

10 Vulgarity

11 Wordplay

12 Other

We believe that most of our humorous tweets can be classified into one of the first 11 categories.

2.1.7 Research on Tweet classification

Much research has been done on classification of Twitter messages into different classes. (Rao, et al. 2010) is a very interesting work that classifies tweets by latent user attributes: gender, age, regional origin and political orientation.

(Pennacchiotti and Popescu 2010) also infer political orientation and ethnicity of tweeters. They used features strange to the tweet content as the profile features, tweeting behavior etc.

(Ramage, et al. 2010) map the tweets' content into dimensions as substance, style, status, and social characteristics. They use latent variable topic models like LDA (Blei, et al. 2003). LDA is an unsupervised model that discovers latent structure in a collection of documents by representing each one as a mixture of latent topics. A topic is represented as distribution of words that co-occur. Much like in this research, their objective was to better find new users and topics on Twitter.

(Silva and Gustavo 2011) manage to classify tweets to 3 authors by examining personal style including the use of punctuation, ‘emoticons’ etc.

(Go and Bhayani 2009) classify tweets by the sentiment they show. This also includes sarcasm detection.

(Davidov, et al. 2010) as mentioned before, classified tweets by trace of sarcasm.

3 Method

3.1 Humor Analysis and Examples

We will now look at some examples of funny tweets, and then review the different types, topics and the way in which the human brain operates in order to “get the joke.” We will also see how computers may try to imitate this:

(1) "And he said unto his brethren, A man shall not poketh another man on facebook for thine is gayeth" \#lostbibleverses

(2) if life gives you lemons, make someone's paper cut really sting

(3) Sitting at a coffee bean and watching someone get arrested at the starbucks across the street. True story.

(4) One of Tigers mistresses got 10 million dollars to keep quiet. I gotta admit I'm really proud of that whore.

(5) There is a new IPod app that translates Jay Leno into funny.

(6) May the 4th be with you...

Example (1) has a hash tag that could help us understand the special stylistic suffixes some words in the sentence bear. These suffixes can be removed by a stemmer.

Search Engine searching for the first part yields more than 2 million hits, since this is a common biblical verse. Humor type **Quotes** can naturally result in a huge amount of search engine results.

The topic is **Facebook**. This can be observed by a computer if we allow it to recognize the named entity “facebook”. Named entities would in many cases serve as the topic.

The word “gay”, will not appear in our lexicon for adult slang but could appear in an insult lexicon (just like “nerd”). So, the computer can identify the imperative tense of the verb “shall not” and together with the insult to infer that this is an **observation** that a very common Facebook action is “gay”. Therefore, the type of this humor would be classified as **observational** and the topic **Facebook**.

Since the word “gay” appears after a copula, we can infer that this is not a regular gay joke where it would usually be an adjective attached to a noun. If it was an “outing” tweet it would not be funny and hence would not have found its way to our dataset in the first place.

For both recognition processes, we require a part of speech tagger and a NE recognizer. We can find these two tools in Alan Ritter's NLP toolkit for Twitter[1] or as we used in this research, as part of NLTK[2], a platform for building Python programs for Natural Language Processing.

Example (2) has no NE or any special lexicon word in it. A Google search of the first part of the sentence, within the quotes, will yield 639,000 results. So we can infer it is of **quote** type. But it is also of type **observation** as it gives a witty comment on life. This can be identified by the words "if" and "life".

Now, why is it funny? The topic is **human weakness**, as described by Mihalcea (2006). We laugh at the manifestation of human misanthropy and the satisfaction in gloating. This relates to the relief theory of humor, as the joke is allowing us to speak about our tabooed and unsocial feelings.

How can the computer understand this? It is a tricky and complex task. We could parse the sentence to find out that the reader is advised to make someone's cut sting, and we could use a semantic ontology or a lexicon to teach the computer that "sting" is a negative experience, which will lead to drawing the correct conclusion. We believe a comprehensive understanding of the sentence is not mandatory, but if necessary, we could use the work of Taylor (2010) as reference.

Example (3) ends with the short sentence "true story," which tells us that this is an **anecdote**. The present progressive tense of the verbs implies the same.

[1] http://www.cs.washington.edu/homes/aritter

[2] http://nltk.org/

To understand this short sentence we need a semantic effort, or a lexicon of such terms that confirm the anecdotal nature of the tweet. The NE “**Starbucks**” could be set as the viable topic.

Example (4) has a proper noun as NE, “**Tigers**”, recognized by its capital initial letter. This is also the topic, and the type is probably **vulgarity**, that can be recognized by the last word in it appearing in a lexicon.

Example (5) is an **insult**, and the topic is the proper name “**Jay Leno**”.

To recognize that this is an insult to Leno, we need to know he is a comedian, and that the tweet suggests that he is not funny. An internet search will discover the former. For the latter, we must understand what “translate something into funny” means. The semantics of the verb and its indirect object that follows the preposition “into” should clarify this. This can be achieved by parsing the tweet, looking up the semantics of “translate” and “comedian” in a semantic ontology, and concluding that Leno is not funny. This is contradictory to his profession and can be viewed as an insult.

Example (6) is a **pun**, or wordplay, in the taxonomy by Hay (1995).

No topic. The pun is based on the phonologic resemblance of “forth” and “force” and the immortal quote from Star Wars. According to Wikipedia, May 4th is actually an official Star Wars day because of this pun, and an internet search of both the original tweet and then the tweet changed back to the original quote (finding the homophone of 4th and replacing it) resulting in very big amount of results can teach our computer what type of tweet this is. Alternatively, with more original phonological puns, phonologic anthologies (which have not been researched thoroughly) can be a proper reference source.

This research is limited to the classification of type alone, and deals with only a few types from the eleven mentioned by Hay (1995) and enumerated in the next section.

3.2 Model Introduction

The model attempts to learn from a seed of labeled tweets how to classify other tweets to whether they belong to a certain humor type or do not.

The classifier builds a model, for each humor type, by learning from a labeled seed of tweets, and uses this model to classify a test set.

For this the model uses many features of many types in an attempt to seize the ones that best discriminate between positive and negative examples, in relation to the humor utilized in them.

The Python auxiliary code enables the tweaking of the minimum rate above which a tweet will be regarded as a positive example.

3.3 Feature Selection

The size of the available labeled data is small, 500 tweets, most of which were negative examples. This was a big obstacle in the experiments, and resulted from the fact that labeling was indeed a long and expensive procedure.

In addition, an automatic method to expand it was not found, due to the complex nature of the labels.

(Tsur et al., 2010) used the adjacency of sarcastic tweets on the web, and the contradiction between Amazon negative star-rating and positive wording in reviews for their automatic seed expansion. Regrettably we could not find a parallel idea to devise such a method in this complex humor case.

Therefore, automatic feature selection gave bad results, and we had to manually choose features for each humor type.

We devised a different feature list for each humor type. The majority of features were common to all types, but some type-specific features were added to each.

Following is our basic feature list:

3.3.1 Stylistic Features

The number of occurrences of the following characters or strings (we preceded some feature names with their identifier, for later reference):
Quote ' “ ',   excl '!',   question '?',   ":-)",   ' ,   '. ',   "... ",   ":",   rbrack "(",   ")"
And the following are specific to **observational** humor:

3.3.2 Lexical Features

The number of occurrences of the following strings:
"when", "when i ",   "when my",   "when you",   "when we", "that moment", "the fact", "if",   "anyone",   "twitter",   " is like",   " are like"
Or one of the strings in this set of first voice pronouns ["i ", "we ", "my ", "our ", " me", " us"," mine"," ours"] (pronounFirst)
As part of preprocessing the tweet we turn it all into non-capital letters in order to recognize all the mentioned strings.
We estimated by reading some of the tweets of this type, that these strings might be of differentiating power, since this kind of tweets is all about personal experiences of the author and many of the above mentioned constructs are used in it.

3.3.3 Morphological Features

The number of gerunds -
Again, we believed that many observational tweets use this form to express present progressive tense of the experience now being observed, or use gerunds to express the action that they ponder upon. E.g. “I love finding money in my pockets after a night of drinking. It's like a gift to sober me…from drunk me.”

For the humor types Jokes, Anecdotes and Fantasy the features counting accurrences of “until” and “would” were added, and the results are shown hereinafter.

3.4 Implementation

The tweets were collected during a few weeks from individual comic tweeters and groups of humor using free software, as described in section 4.1.

Pre-processing was performed on the tweet data table to prepare them for the labeling step and then a crowd sourcing online tool was hired for this task. 505 tweets were labeled with a 5 rank rating for whether they are funny, and then how relevant they are, or how strong they can be classified, to each of the 11 humor types defined to the labelers.

Our python code had more pre-processing performed on the tweets to normalize them for the feature examining, e.g. all capital letters were changed to regular ones. It also prepared the tweets to be input to the parsing algorithm we used to find gerunds.

The selected features were then examined for each tweet and their values recorded.

The machine learning software WEKA was then used in 10-fold experiments on a few humor types and a few classifying algorithms, to build a model for the differentiation of tweets that pertain to a certain humor type from those who do not.

4 Experiments

4.1 Data Description

Our task is to categorize the different humorous tweets.

A little about Twitter:

Twitter is a popular micro-blogging service with more than 200 million messages (tweets) sent daily. The tweet length is restricted to 140 characters. Users can subscribe to get all the tweets of a certain user, and are hence called followers of this user, but the tweets are also publically available, and can be read by anyone. They may be read on the Twitter website, on many other sites, and through Twitter API, an interface that allows access to a great amount of tweet and user attributes.

Aside from text, tweets often include URL addresses, references to other Twitter users (appear as "@<user>") or content tags (called hashtags and appear "#<tag>"). These tags are not taken from a set list but can be invented by the tweeter. They tend to be more generic since they are used in Twitter's search engine to find tweets containing the tag, so users try not to over-invent.

Our humorous tweet dataset is derived from Twitter´s comic tweeters like @BestAt, @comedyEpic, @FunnyJokeBook, @TheComedyJokes. These tweeters are all from USA. Since Twitter does not allow access to old tweets, an online capture procedure of tweet stream was required, performed by the twitter-to-pdf software.[3]

The tweets collected were then processed to eliminate the ones containing URLs, because this research only deals with textual humor, and the ones shorter than 15 characters, since it would contain too little information for our purposes.

4.2 TRAINING DATA ACQUISITION

Categorizing humor is of course very complex, due to the fuzzy nature of the taxonomy and the subjectivity of this task. One tweet can belong to more than one humor type, and two people may have different ideas on those types, as we could see in the data.

[3] http://sourceforge.net/projects/twitter-to-pdf/

Nevertheless, the only way to achieve a gold standard for such classification is through human annotation, which can be accomplished through the use of a mechanical Turk. Thus, labeling of the whole dataset was called for, since no automatic method could be devised to label the humor type or to collect new tweets of the same type based on a labeled seed. Additionally, human labeling is expensive, so the size of the dataset was naturally limited.

Hence, 500 of the collected tweets were labeled with the types of humor they contained. For this task a crowd sourcing (or Mechanical Turk) service was hired.

This service offers a tailored solution for sentiment analysis tasks, called Senti.

The upload of the data and the preparation of the questionnaire and instructions given to labelers are all automatic through their website[4].

The labeling procedure included mainly annotators from all over the USA, so they would better understand the humor, when suitable cultural background is a must.

Each tweet was labeled by more than one labeler, each first deciding if the tweet was funny at all, and if so, deciding if each of the 11 humor types is relevant to that tweet, i.e. the tweet can be classified to that type of humor. If relevance was marked, then a grade between 1 and 5 is also required for the strength of the sentiment, e.g. to what extent does the tweet include anecdotal humor.

Following are the eleven topics and a short explanation of each as given to the labelers as instructions:

- **Anecdotes**. A story which the author perceives to be amusing, about the experience or actions of himself or someone they know. e.g. "Sitting at a coffee

[4] http://crowdflower.com/

bean and watching someone get arrested at the Starbucks across the street. True story. "

- **Fantasy**. Humorous, imaginary scenarios or events. E.g. "If weed becomes legal, I can't wait to see the commercials." or "I will not be impressed with technology until I can download food from the internet."
- **Insult** is a remark that puts someone down, or ascribes a negative characteristic to him. e.g. "There is a new IPod app that translates Jay Leno into funny"
- **Irony or sarcasm**. The author in his words is implying the opposite, or something with a clearly different meaning, e.g. "Oh 50+ likes on your picture? I'm sure you were well dressed. " OR blaming others with his fault (Pot calling the kettle black)
- **Joke**. A chunk of humor whose basic form has been memorized. They often have a standardized form (narrative, Q&A...). E.g. "Mom: ‘How was school today?’ Me: ‘Uhh, we had a surprise test today.’ Mom: ‘And?’ Me: ‘...I was surprised.’ "
- **Observation**. Comments about the environment, the events occurring at the time or life. The author is making an observation about something funny, or making a witty observation, e.g. "That awkward moment when X adds you on Facebook."
- **Quote**. A line taken from a TV show or a movie, usually a comedy, e.g. "May the 4th be with you" (a quote and a wordplay) or "What happens in Vegas stays in Vegas. What happens on Twitter, wow it sure gets around fast! " (quote and paraphrasing)
- **Self Deprecation**. The author directs an insult to himself, e.g. "Taking dumb risks is in my DNA code."

- **Vulgarity**. The author is breaking some sort of taboo in certain environments, usually about sex and toilet. E.g. "Taking a shit so intense you gotta take your shirt off "
- **Wordplay**. Any humorous statement in which the humor derives from meanings, sounds or ambiguity of words, e.g. "Make the little things count. Teach midgets math" or "If you rearrange letters in Mother-in-law, you get Woman Hitler. Coincidence? I think not."
- **Ethnic** humor. The text attributes certain characteristics to a certain ethnic group. e.g. "I'm not racist, because racism is a crime. And crime is for black people." (ironic and ethnic) or "How does Moses make his tea? Hebrews it. Then he sells it for prophet." (wordplay and ethnic)

The results of all labelers were then aggregated using the method explained in the appendix to give a single set of results for each tweet: humor relevance and rating, 1-11 type relevance and rating.

The crowd sourcing service's friendly and interactive graphical user interface, as seen in figure 1, allows easy examination of charts and the data collected.

The labeled data was then preprocessed to delete tweets labeled as not funny or if none of the 11 topics was found relevant to them.

The SASI algorithm from (Tsur et al. 2010) was carefully examined in this work however we could not base our research on theirs, because we thought that humor types have little to do with surface patterns, and more with the data content. In addition, we had a considerably smaller size of dataset resulting from the expensive human annotation and the inexistence of an automatic way to expand it. Tweets of a certain humor type are not necessarily tweeted adjacent to others of the same type. A smaller dataset is not likely to extract pattern features successfully.

In addition we could not come up with a baseline for evaluation other than ZeroR- "pick the majority class" since we had no additional information about the type of the humor exposed in the tweet, for a heuristic baseline. Due to the small amount of positive example of every type in our set, this choice of baseline should be taken into consideration when comparing the results of our experiments.

We could not use a semi-supervised approach as well, in spite of the expensive annotation, because of the lack of seed expansion ability, as far as we could reach.

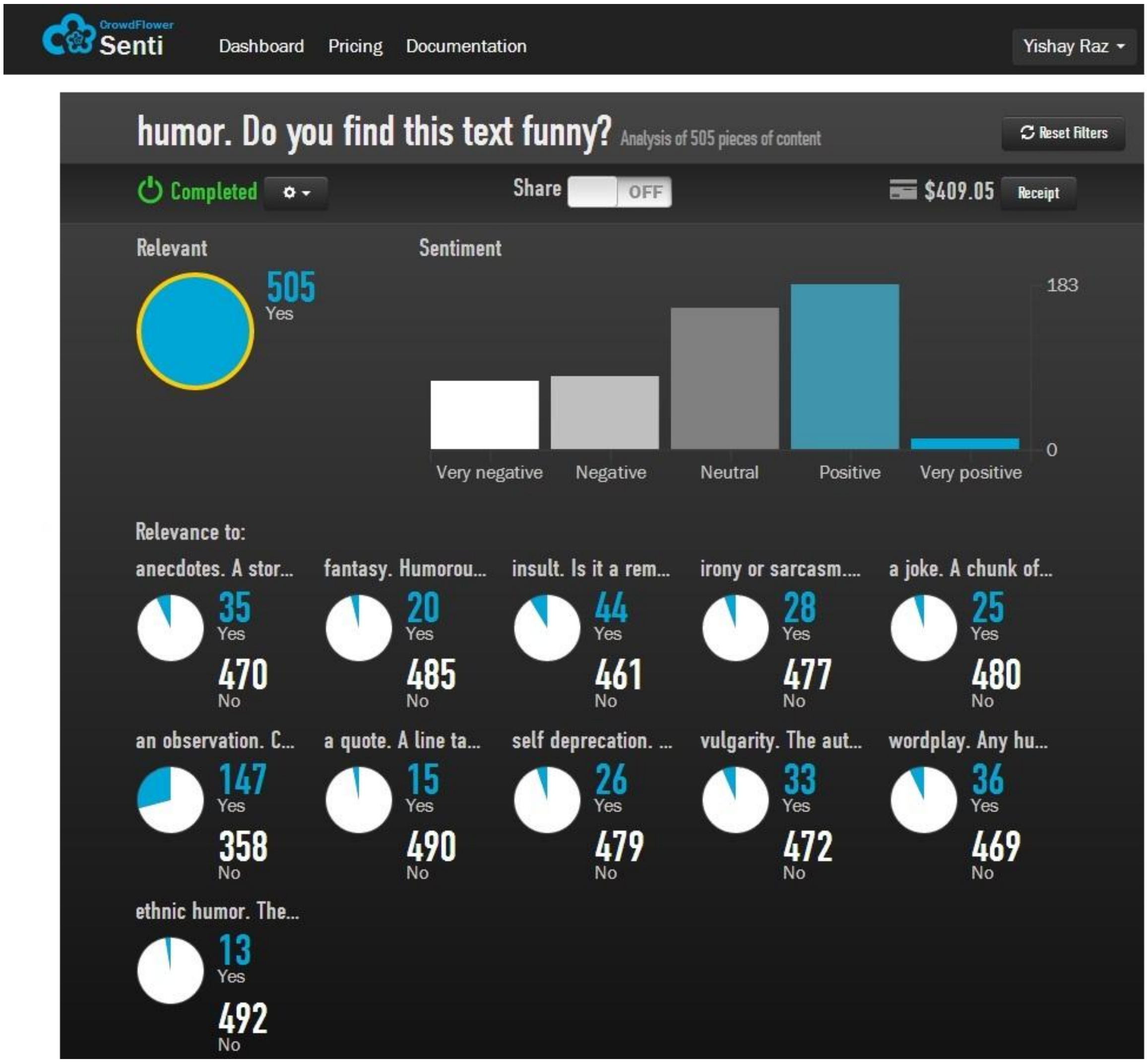


Figure 1 – Screen print from the Crowd Flower dashboard showing the labeling results

All tweets, including those given a bad relevance rating for humorous, were used in our experiment for training and testing.

The reason is that we believe that humor is subjective, and low rated tweets could be considered funnier by some people.

On the other hand, when a topic was given a poor relevance grade, we gave the possibility in the software for it to be disregarded. The rating threshold above which tweets were considered to be of a certain topic could be controlled.

We ran this experiment both for a threshold of 0 (any relevance to the topic) and of 3.

The results of both experiments, which are shown hereinafter in the following tables, are different due to the big difference in dataset size and proportion of the classes in it.

Qualifying tweets were then used to train and test a classifier.

10-fold stratified cross-validation was used to divide the hundreds of tweets into mutually exclusive training and testing sets, the former in the size of nine tenths of the initial set, and the latter, the remaining one tenth.

Every experiment was performed on one humor type, i.e. training and testing was run on a dataset of tweets each labeled by a Boolean referring to whether the tweet has a certain type of humor (e.g. observational) in it.

4.3 EXPERIMENTS ON HUMOR CLASSIFICATION

4.3.1 The classifiers

The classification experiments, including both training and testing were done using WEKA software package[5] (University of Waikato, New Zealand).

We ran the classification experiments with the following four algorithms:

[5] http://www.cs.waikato.ac.nz/ml/weka/

4.3.1.1 Naïve Bayes

This is the basic example of a probabilistic classifier. It applies a slightly modified version of Bayes' theorem and is called Naïve because of its assumption that the variables (features) are independent of one another.

This algorithm estimates the parameters for classification:

The first group is the class priors, which is the chance of a tweet tested to belong to a certain class, or in other words this class' proportion in the dataset.

The second group is the feature probability distribution, which are the different probabilities of finding a specific feature f in a tweet, given that the tweet belongs to a specific class. We then use these parameters to classify tested tweets by this formula:

$$p(C \mid F_1, ..., F_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(F_i \mid C)$$

4.3.1.2 C4.5 (J48 implementation)

A decision tree building algorithm (Quinlan, 1993)

It uses the concept of information entropy: at each node of the tree, C4.5 chooses one feature of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The feature with the highest information gain (difference in entropy) is chosen.

*4.3.1.3 K**

(Cleary and Trigg, 1995) developed this algorithm. It is an instance-based classifier which classifies a test instance according to the class to which the instance that resembles it in the training set belongs. The resemblance is determined by a similarity function. This function is the entropic distance measure.

0-R - is the trivial classifier that classifies all items to the class most seen in training, i.e. the mean (for a numeric class) or the mode (for a nominal class as ours).

This classifier is used as our baseline, and all other results will be compared to its results.

4.3.2 Feature optimization

Now, we use the wrapper method through WEKA to select the minimal set of features that gives the best results. We use the CfsSubsetEval attribute evaluator and the BestFirst search method.

Since we have already proven previously the feasibility of classifying unseen tweets, we can now use the whole dataset for feature selection, in order to prepare a better and sleeker feature set for future experiments and new datasets.

The simplest selection algorithms test all subsets of features, and choose the one that maximizes accuracy. They differ on the way they estimate this accuracy.

Wrapper methods use a predictive model to score feature subsets. Each new subset is used to train a model, which is tested on a hold-out set. Counting the number of mistakes made on that hold-out set gives the score for that subset. As wrapper methods train a new model for each subset, they are very computationally intensive, but usually provide the best performing feature set for that particular type of model.

The results do show that some of the humor-type specific features that we devised are good enough to be included in the selected set.

Using J48 we got six selected attributes: 1,3,6,7,16,23 which are: Quote, question, when, wheni, dot, pronounFirst. Figure 3 shows the pruned decision tree built with these features.

```
when <= 0: 0 (329.0/103.0) [2]
when > 0
|   when <= 1
|   |   wheni <= 0
|   |   |   dot <= 1
|   |   |   |   question <= 0
|   |   |   |   |   quote <= 0: 1 (29.0/1.0) [3]

|   |   |   |   |   quote > 0
|   |   |   |   |   |   pronounFirst <= 1: 0 (2.0)
|   |   |   |   |   |   pronounFirst > 1: 1 (5.0/1.0)
|   |   |   |   question > 0
|   |   |   |   |   pronounFirst <= 0: 1 (4.0/1.0)
|   |   |   |   |   pronounFirst > 0: 0 (4.0)
|   |   |   dot > 1: 0 (7.0/2.0)
|   |   wheni > 0
|   |   |   pronounFirst <= 4: 0 (7.0/1.0) [1]
|   |   |   pronounFirst > 4: 1 (3.0/1.0)
|   when > 1: 1 (4.0)
```

Figure 2 : J48 pruned tree after feature selection

Using Bayes four attributes were selected: 1,2,6,15 which are Quote, excl, when, twitter

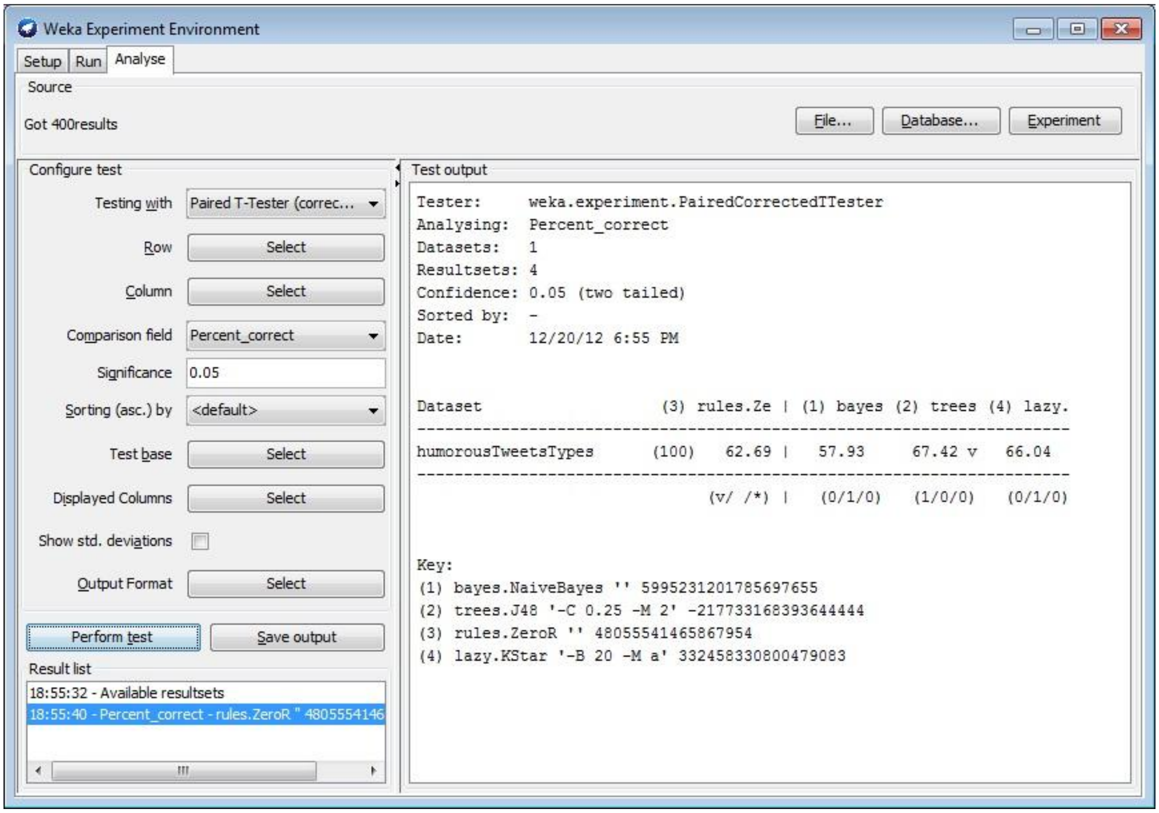
Using Kstar fourteen: 1,2,3,6,7,9,12,13,15,16,18,20,23,24 which are:

Quote, excl, question, when, wheni, whenyou, thefact, if, twitter, dot, colon, rbrack, pronounFirs, Gerunds.

Quote and when were chosen in all three algorithms.

4.3.3 The results

In figure 2 and table 2 we show the results of a ten-fold experiment of classifying tweets to type observational humor. We see that the DT algorithm gives results that are statistically better (confidence 0.05 two tailed) than the baseline: ZeroR classifier. The percent of correctly classified tweets is higher in 5 points and the precision is higher in 0.08 points. This result is even more interesting when we keep in mind that almost 63% of the tweets are not anecdotes, i.e. ZeroR has 63% of accuracy and not the usual around 50% as would have been the case in a more evenly distributed dataset.



Got 400results

Testing with: Paired T-Tester (correc...

Comparison field: Percent_correct

Significance: 0.05

Sorting (asc.) by: <default>

Show std. deviations □

18:55:32 - Available resultsets
18:55:40 - Percent_correct - rules.ZeroR '' 4805554146

```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       12/20/12 6:55 PM


Dataset                   (3) rules.Ze | (1) bayes (2) trees (4) lazy.
----------------------------------------------------------------------
humorousTweetsTypes      (100)   62.69 |   57.93     67.42 v   66.04
----------------------------------------------------------------------
                                (v/ /*) |   (0/1/0)   (1/0/0)   (0/1/0)


Key:
(1) bayes.NaiveBayes '' 5995231201785697655
(2) trees.J48 '-C 0.25 -M 2' -217733168393644444
(3) rules.ZeroR '' 48055541465867954
(4) lazy.KStar '-B 20 -M a' 332458330800479083
```

Figure 3: Comparison of the 3 algorithms' results in WEKA, rating threshold=0, 394 tweets

Humor type	(1) bayes.Na	(2) trees	(4) lazy.
Anecdotes	√		
Fantasy	√		
Insult		√	
Irony			√
Jokes		√	
Observational		√	
Quote		√	
Role play	√		
Self deprecation			√
Vulgarity		√	
Wordplay			√

Table 1: The algorithm that achieved best results (F-score) for each humor category (rating threshold=0)

Dataset	(3) rules	(1) bayes.Na	(2) trees	(4) lazy.
Precision	0.62	0.70 v	0.67	0.67
Recall	1.00	0.73 *	0.94	0.74 *
F-score	0.76	0.70	0.77	0.69

Table 2: The 3 algorithms' results, type: observational, rating threshold=0, 394 tweets

We see that when the rating threshold picked is higher, we have significantly less tweets in the dataset, and the rate of the tweets tagged as including observational humor is 10%. Therefore, the algorithms do not perform well, as they can classify, like ZeroR, all tweets to the majority class, and still show good results. These results are shown in table 3.

Dataset	(1) bayes.Na
Precision	0.68
Recall	0.68
F-score	0.65

Table 3: The NB algorithm results, type: observational, rating threshold=0, **4 selected features**

Dataset	(3) rules	(1) bayes.Na	(2) trees	(4) lazy.
Percent correctly classified	91.82	88.00	91.82	89.73
Precision	0.92	0.93	0.92	0.92

Table 4: The 3 algorithms' results, type: observational, rating threshold=3, 110 tweets

Unfortunately, the tweets tagged with all other humor types in our dataset constitute less than 9% of it, and hence adequate learning by the classifiers was not achieved, and usually the results did not show any significant advantage over the baseline, as can be seen in the tables below.

Nevertheless, examining the features selected by the selection algorithm showed interesting results: The only feature chosen out of 27 to classify fantasy was the number of occurrences of the word "would". The two chosen for anecdotes were "when I" and gerunds, and the two for jokes were quotation and question marks. Following are examples from our dataset of one tweet for each type that will show the relevance of the chosen features to the corresponding humor type:

- Fantasy: A world without women **would** be a pain in the ass
- Anecdote: My family looks at me strangely **when I** randomly start **laughing** at the computer.
- Joke: **"**Hey, what's up**?"** -Gas prices.- "You know what I mean, like what's crackin?" -Nutshells.- "Really? Fine. What's poppin?" -Corn.-

As shown in table 2 above, classifying with fewer selected features affects the results to the worse, but a compromise should be reached regarding this trade-off: a good enough precision with a smaller set of features, especially when the feature list grows bigger, and introduces time and resource restrictions, and also for the sake of simpler and more understandable feature list and DTs, as shown in the examples above.

Dataset	(1) bayes.Na	(2) trees	(3) rules	(4) lazy.
Precision	0.92	0.93	0.92	0.92
Recall	1.00	0.94 *	1.00	0.98
F-score	0.96	0.93	0.96	0.95

Table 4: The 3 algorithms' results, type: anecdote, rating threshold=3

Dataset	(3) rules.Z	(1) bayes	(2) trees	(4) lazy.
Precision	0.94	0.98 v	0.96 v	0.95 v
Recall	1.00	0.95 *	0.97 *	0.99 *
F-score	0.97	0.97	0.96	0.97

Table 5: The 3 algorithms' results, type: jokes, rating threshold=0

The Rules algorithm above refers to ZeroR: a 0-R classifier predicts the mean (for a numeric class) or the mode (for a nominal class). In our case, since the mode is 0, i.e. tweet contains no **observational** humor, all tweets will be classified as such, and the percent of correctly classified tweets will be almost 63.

The annotation "v" or "*" indicates that a specific result is statistically better (v) or worse (*) than the baseline scheme at the significance level specified (currently 0.05).

4.4 Comparison with Other Methods

No other methods or research were found about classifying humor types.

In the related work section, we revised a number of researches and methods that resemble this work's subject. Among them, the most similar are the work of Tsur and Davidov on the recognition of sarcasm, and that of Mihalcea on humor recognition.

In this work we tried to use the ideas and methods described in these past researches and implement them as much as we could. Unfortunately, many of them resulted inappropriate for this work because of the special character of the data investigated here.

4.5 ERROR ANALYSIS

The nature of classifiers, always and especially when classifying humor is pure statistical. The intelligence required to classify a joke to its humor type is too complex to be taught to a computer and the heuristics learnt in the classifier's construction, while learning the labeled examples, are rules of thumb and not sacred. Therefore, errors are inevitable and abundant. We will look at a few of them, relating to the false positives of figure2. The relevant node in the tree figure is marked with the number of the example in red and superscript (e.g. ^{1}):

1. "When I was a kid, Pussy meant CAT, Sex meant GENDER, Dick was a NAME, and Bang was a SOUND."

   This tweet is falsely classified as non-observational. The crowd gave it a 3 of 5 as observation-relevance-score. The reason for the false classification is, as you can see in the tree, the fact that "when I" appears in it but with less than 5 first person pronouns (only "I"). This is the only tweet badly classified at this node, with 7 correctly identified.
2. "My life: Wake up, survive, sleep" is an example of the 103 tweets falsely recognized as non-observational because they contain no "when".

3. "I miss being a kid, when the biggest decision of your day was picking a crayon" was the only tweet classified falsely as non-observational at this decision node, while 29 others were classified truly because they all include 1 "when" not followed by an "I", less than 2 dots, no question marks and no quotes.

5 Conclusion

We have shown in the scope of this thesis that machine learning techniques are relevant for the very challenging task of humor classification. The experiments have shown that common classifiers as NB and DT can give better and statistically significant results and manage to classify tweets by whether they belong to a certain humor type, whose features it learned, or they do not. Despite the limited resources, a proof of concept is achieved, and this field could and should be the subject of future endeavors.

5.1 FUTURE WORK

The features used in this work were basic, and the research on the following more sophisticated ones is bound to produce a bigger success:

Syntactic Features

Transitivity of the verb

Syntactic ambiguity

Pattern-based Features

Patterns including high-frequency and content words as described in the algorithm in (Davidov and Rappoport 2006)

Lexical Features

Lexicon words as adult slang, ethnic groups, insults etc.

Existence of NEs (like Facebook and Starbucks)

Semantic meaning of the verb and its objects (“make someone’s cut sting”)

Lexical ambiguity

Morphological Features

The tense of the verbs in the tweet (we only used gerunds)

Special word morphology (like the biblical “eth” suffix in our example (1))

Phonological Features

Existence of a word that appears on a homophones list (which could help with pun recognition)

Pragmatic Features

The amount of results obtained from a search engine query of the tweet (good for quotes)

Stylistic Features

The existence of smiley characters etc.

Furthermore, a much bigger dataset should be applied, and maybe an automatic way to acquire tweets, knowing their type, can be devised. Otherwise, the human annotation will always limit out data for being expensive.

In addition to the resources, the objectives of the research can also be elaborated.

The topic of a tweet can also be retrieved either from the text NE or from automatically retrieved features when it does not appear as a NE in the tweet.

Further research can be done to classify the tweeters of the humorous tweets based on attributes of gender, age, location, etc. This could be achieved using the type and the topic of the tweets as additional features to semi-supervised classifiers.

This idea was inspired by related work of Pennacchiotti and Popescu (2011) that found a correlation between humor types produced by different groups of gender, ethnic groups etc. and other work mentioned in the Related Work section.

Another aspect to be researched is the psychological state of the tweeter as reflected in his tweets. In related work we have seen that humor is used to break taboos and "release steam" and this could mean that, together with cynicism, it can result in people being more sincere or closer to their feelings and expose deeper more subconscious thoughts in their humor. This could serve as a great tool in locating people suffering from a harmful psychological state, whether they are suicidal or plan a massacre. Of course, this field is a very sensitive one, but we feel that the research possibilities in it are very intriguing and immense.

Appendix

How we chose the label when labelers were not unanimous

As mentioned above, each tweet in the dataset was tagged by more than one tagger. This was of great importance due to the very subjective nature of humor and of the 12 humor types given to the taggers to choose from.

CrowdFlower offers different types of aggregation and we chose the following:

`agg`

> This outputs the answer with the highest agreement weighted by worker trust *(confidence)*. A numerical confidence value between 0 & 1 is also returned.

How we choose the answer and calculate confidence (aggregation=”agg”)

1. Find the sum of trust for each response
    a. Sum of trust(beef) = 4.4703
    b. Sum of trust(chicken) = 1.8571
    c. Sum of trust(veggie) = 0.9231

The answer with the maximum sum of trust is output.

2. Find the sum of trust for all responses
    a. Sum of trust(all) = 7.2505

3. Divide the max of (1) by (2) to find confidence for that unit
    a. Confidence = 0.6166

_unit_id	_id	_tainted	_channel	_trust	_worker_id	_ip	best_burrito
81854629	139411475	FALSE	Prodege	0.9231	2981776	24.56.163.117	veggie
81854629	138189987	FALSE	Prodege	0.8121	2912496	184.17.34.86	beef
81854629	139481118	FALSE	Prodege	0.9333	2982300	76.188.0.26	beef
81854629	139542120	FALSE	Gambit	0.8571	993194	96.252.78.158	beef
81854629	139743646	FALSE	Amt	0.963	1680385	76.118.74.186	beef
81854629	139847442	FALSE	Cotter	0.9048	607699	12.96.96.5	beef
81854629	139929185	FALSE	Prodege	0.8571	2866321	76.90.124.214	chicken
81854629	138072064	FALSE	Prodege	1	2542397	75.172.49.134	chicken

In this case, the aggregate csv for this unit would be the following:

_unit_id	_golden	_trusted_judgments	best_burrito	best_burrito:confidence
81854629	FALSE	8	beef	0.6166

Glossary

API - Application programming interface

CML – CrowdFlower Markup Language 's

DT – Decision Tree

LDA – Latent Dirichlet Allocation

NB – Naïve Bayesian

NE - Named Entity

References

Attardo, S., Raskin, V. 1991. Script theory revisited: Joke similarity and joke representation model. Humor: International Journal of Humor Research 4, 3-4.

Bird, Steven, Edward Loper and Ewan Klein (2009). Natural Language Processing with Python. O'Reilly Media Inc.

Cheng, Z., Caverlee, J., Lee, K. 2010. You Are Where You Tweet: A Content Based Approach to Geo-locating Twitter Users. Proceeding of the ACL conference 2010

Cleary, J.G. and Trigg, L.E., 1995. K*: An Instance-based Learner Using an Entropic Distance Measure. In: 12th International Conference on Machine Learning, 108-114.

Davidov, Dmitry, and Oren Tsur, 'Enhanced Sentiment Learning Using Twitter Hashtags and Smileys', 2007

Davidov, D., and Tsur, O. 2010. Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon Computational Linguistics, July, 107-116.

Dias de Oliveira Santos, V., 2011. Automatic Essay Scoring: Machine Learning Meets Applied Linguistics. Master Thesis.

Freud, S. 1905. Der Witz und seine Beziehung zum Unbewussten

Freud, S. 1960. Jokes and their relation to the unconscious International Journal of Psychoanalysis 9

Go, Alec and Bhayani, Richa, 'Exploiting the Unique Characteristics of Tweets for Sentiment Analysis ∗', Read, 2009

Hay, J. 1995. Gender and Humour: Beyond a Joke. Master thesis.

Hay, J, 'Functions of Humor in the Conversations of Men and Women', Journal of Pragmatics, 32 (2000), 709–742 <doi:10.1016/S0378-2166(99)00069-7>

Hobbes, T. 1840. Human Nature in English Works. Molesworth.

Hurley, M.M., Dennett, D.C. and Adams Jr., R.B. 2011. Inside Jokes - Using Humor to Reverse-Engineer the Mind. MIT

Mihalcea, R. 2006. Learning to laugh automatically: Computational models for humor recognition. Computational Intelligence, 222.

Mihalcea, R. and Strapparava, C. 2005. Making Computers Laugh: Investigations in Automatic Humor Recognition. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing

Minsky, M. 1980. Jokes and the logic of the cognitive unconscious. Tech. rep., MIT Artificial Intelligence Laboratory.

Nevo, O., ‘Appreciation and Production of Humor as an Expression of Aggression: A Study of Jews and Arabs in Israel’, Journal of Cross-Cultural Psychology, 15 (1984), 181–198 <doi:10.1177/0022002184015002006>

Pennacchiotti, M. and Popescu, A. 2011. Democrats , Republicans and Starbucks Afficionados: User Classification in Twitter . Statistics, 430-438.

Pennacchiotti, M. and Popescu, A. 2010. ‘A Machine Learning Approach to Twitter User Classification’, Artificial Intelligence, 281–288

Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

Silva, Rui Sousa, and Gustavo Laboreiro, ‘Automatic Authorship Analysis of Micro-Blogging Messages’, ReCALL, 2011, 161–168

Ramage, Daniel, and Susan Dumais, ‘Characterizing Microblogs with Topic Models’, International AAAI Conference on, 2010

Rao, D., Yarowsky, D., Shreevats, A. and Gupta, M. 2010. Classifying Latent User Attributes in Twitter. Science, 37-44.

Raskin, V. 1985. Semantic Mechanisms of Humor. Kluwer Academic Publications

Solomon, R. 2002. Ethics and Values in the Information Age. Wadsworth.

States, Western, Folklore Society, and Western Folklore, ‘Jewish Humor Strikes Again : The Outburst of Humor in Israel During the Gulf War’, Society, 53 (1994), 125–145

Taylor, J. M. 2010. Ontology-based view of natural language meaning: the case of humor detection. Journal of Ambient Intelligence and Humanized Computing, 13, 221-234.

Tsur, Oren, and D Davidov, ‘ICWSM – A Great Catchy Name : Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews’, Proceedings of the Fourth International, 2010, 162–169

Ziv A. 1988. National Styles of Humor. Greenwood Press, Inc.