Searching for Puns: Towards Identification of Humour on Twitter

Morgan Harvey
Faculty of Informatics
University of Lugano (USI)
morgan.harvey@usi.ch

ABSTRACT

Research has shown that casual leisure search - when users are searching for entertainment purposes and with no fixed information need to fulfil - is becoming an increasingly common type of online task. A frequently occurring type of casual leisure search involves users looking for something humorous to help pass the time, such as a short gag or pun. An almost unlimited source of such material is Twitter, a service which allows millions of users to post short messages (known as *tweets*) to their followers. Despite the potential offered by Twitter, it is often very difficult to extract good quality content from the huge number of nonsense messages posted. In this work we conduct a small user study (n=8) to try to learn whether people agree on what a humorous tweet is and discuss ways in which we could learn how to automatically identify funny jokes, gags and puns posted on Twitter.

1. INTRODUCTION AND MOTIVATION

Although the traditional view of search on the web is that the user has some kind of *information need* in mind that they wish to fulfil, recent work has shown that an increasingly large number of queries are for entertainment and hedonistic purposes [**?**, **?**], so-called "casual leisure search". People are spending more and more time online, browsing around the Internet with no specific goal in mind and simply looking to alleviate boredom, to find something fun to do or to pass the time. A common way of spending time on the web in recent years has been to browse socially-generated content streams such as those available on Facebook and Twitter. This is becoming so popular that by May of 2013 almost three quarters of all online U.S. adults were using social networking sites with over a quarter of these having their own Twitter accounts [**?**].

Twitter allows users to post short "tweets" - messages of up to 140 characters - which are immediately shared with everyone who follows the posting user. Twitter is used for a wide variety of different reasons, including in a number of casual leisure situations such as looking for music and images as well as humorous or entertaining content within the tweet itself [**?**]. Unfortunately the huge number of pointless tweets often flood the relevant tweets, making the discovery of such entertaining content on Twitter very difficult [**?**, **?**]. In this work we focus on the problem of identifying gags, jokes and puns on Twitter - which are commonly retweeted by users, suggesting that they like them [**?**] - and conduct a small user study (n=8) to determine whether people agree on what a humorous tweet is.

2. FINDING AND EVALUATING TWEETS

Before we subject users to the somewhat onerous task of classifying short messages based on how humorous they are, we need to ensure that we have a reasonable number of candidate tweets and that a reasonable proportion of these contain puns, jokes or gags. To do this we rely on Twitter's lists system which allows users to create and curate groups of other Twitter users, often based around a single topic or theme, which other users can then subscribe to. An initial search of Twitter returned three candidate lists containing users who often posts funny content. We then used the Twitter API [1] to expand our set of candidates by downloading the profiles of all the users who are members of the existing three lists. For each of these users we then added to our set of candidate lists all of the lists they were subscribed to which contained one of the following words in their descriptions or titles: "gag", "joke", "pun", "humour", "humorous" and "funny," yielding a final total of 179 candidate lists. The lists had an average of 173 members and 72 subscribers and were curated by 25 different Twitter users.

For each of these lists we downloaded the last 200 tweets posted to each of them (ignoring all duplicates), resulting in a final total of 30,072 candidate funny tweets. Finally we chose two subsets of these tweets of size 200 at random for our participants to evaluate. We recruited 8 participants for our study, ensuring that there was an equal balance between native speakers of English and non-native speakers (i.e. 4 of each). Three participants were in the age bracket 21-30, two were between 31 and 40 and the remaining three were aged between 51 and 70. Three participants were female. The participants were separated into two groups, one for each set of 200 tweets, with the stipulation that the equal balance of native and non-native speakers was maintained.

Each participant was sent a personalised link to a web form which allowed them to evaluate the candidate tweets in terms of how humorous they were. Due to the very subjective nature of humour, users were given the option to choose between two grades of humorousness for tweets they found funny: "funny" and "vaguely humorous." Users could also indicate that the tweet may have been intended as humour but that they personally did not find it funny. Some of the tweets (39 out of 400) also contained an image which was displayed below the textual content. An example screenshot demonstrating the tweet evaluation form is shown in

[1]Twitter REST API version 1.1:
https://dev.twitter.com/docs/api/1.1

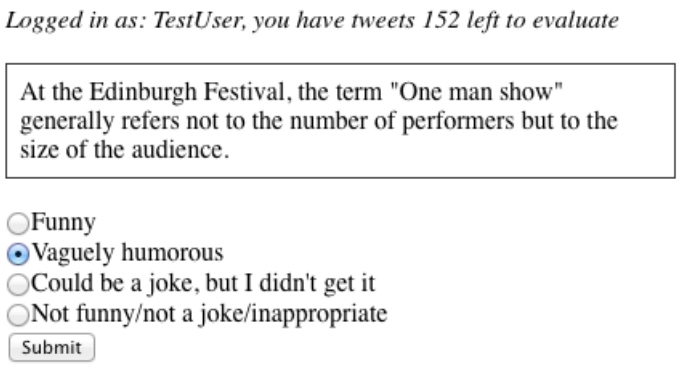

Figure 1: Screenshot of tweet evaluation form.

figure **??**. An example of a tweet universally agreed to be humorous is "I fear change.Which is probably why I lost my job at the bank."

3. RESULTS

In order to evaluate the level of agreement between participants in our study we used Fleiss' Kappa which calculates the degree of agreement in classification over that which would be expected by chance. The evaluations from group one returned a Fleiss' Kappa score of 0.235 (z=12.3, $p \ll 0.01$), indicating a "fair agreement" according to Landis et al. [**?**]. Perhaps unsurprisingly, the level of accordance between the two English native speakers was even higher: Fleiss' Kappa 0.263. The level of agreement between participants in group two was somewhat lower (Fleiss' Kappa of 0.137, z=5.2, $p \ll 0.01$). However, this still indicates the the level of agreement is much higher than would be expected by chance, indicating that the participants did agree in a large number of cases. It is interesting to note that agreement between the two female participants in this group was much higher (0.258).

Although the results above already suggest that people can to a certain extent agree on what is humorous we wanted to see whether the existence of two grades of humour in the form had introduced too much subjectivity. To do this we conflated the categories "funny" and "vaguely humorous" into a single evaluation class and re-analysed the data. Doing so yielded much higher levels of agreement for both groups: 0.37 for group one and 0.308 for group two. Analysing the results from group one in more detail we find that there are 17 tweets that participants unanimously agreed were funny and 89 they all agreed were not funny. If we take a majority voting system (i.e. a tweet is deemed funny if three or more participants agree) then we find 38 funny tweets and 129 unfunny ones, with only 33 being ambiguous (32,133 and 35 respectively for group two). In general, native speakers tended to agree more often with each other on the classification of a tweet than non-native speakers and there were several (n=13/200) cases where both native speakers though a tweet was funny but the non-native speakers did not agree. Many of these cases were one-liners with some level of word play, for example: "I'm starting a bungalow security company if anyone is looking to get in on the ground floor." In comparison there were no instances where the non-native speakers both found a tweet humorous and the native speakers didn't.

4. DISCUSSION

In this short paper we have begun to address the problem of automatically identifying humorous tweets which people may wish to read. A necessary first step in achieving this is to find out if people can even agree on what constitutes a humorous tweet in the first place. To do this we conducted a small user study in which we asked 8 participants to rate the humorousness of tweets based on a simple 4-point scale. By assessing the level of agreement between participants using Fleiss' Kappa, we found that the problem of assessing humour is not quite as subjective as one might think as a fair level of agreement was observed, particularly between native speakers. By using a majority voting scheme we were able to identify 70 funny tweets and 262 unfunny ones which could be used as relevance judgements for building classification models.

To build such models, it would be necessary to generate features of tweets that carry some amount of discriminative power and then learn a model based on these features which maximises the likelihood of the positive relevance judgements (i.e. the funny tweets) being assigned a positive class label. A large number of classification methods would be suitable for this purpose, including logistic regression models, support vector machines and decision trees. Features could be derived from the content of the tweets including the use of specific humorous words or certain parts of speech. Features could also be extracted from the basic statistics of each tweet, for example the number of times it was retweeted and favourited or the number of humour-related lists its author belongs to. We leave these possibilities for future work.

5. REFERENCES

[1] P. Analytics. Twitter study–august 2009. *San Antonio, TX: Pear Analytics. Available at: www. pearanalytics. com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009. pdf*, 2009.

[2] J. Brenner and A. Smith. 72*http://pewinternet.org/Reports/2013/social-networking-sites/Findings.aspx last accessed on 21.11.2013*, 2013.

[3] D. Elsweiler and M. Harvey. Engaging and maintaining a sense of being informed: Understanding the tasks motivating twitter search. *JASIST*, 2014.

[4] D. Elsweiler, M. Wilson, and M. Harvey. Searching4fun. In *ECIR workshop*, 2012.

[5] M. Krieger and D. Ahn. Tweetmotif: exploratory search and topic summarization for twitter. In *In Proc. of AAAI Conference on Weblogs and Social*, 2010.

[6] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[7] M. L. Wilson and D. Elsweiler. Casual-leisure searching: the exploratory search scenarios that break our current models. In *4th International Workshop on Human-Computer Interaction and Information Retrieval*, August 2010. New Brunswick, NJ.