Linking Humour to Blogs Analysis: Affective Traits in Posts

Antonio Reyes and Paolo Rosso
Natural Language Engineering Lab - ELiRF
Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, Spain
{areyes,prosso}@dsic.upv.es

Abstract. The Web 2.0 technologies are gradually changing our intercommunication habits. Means such as forums, blogs, wikis, social networks, etc., are a clear sample about the importance of these tools for expressing and sharing contents, ideas, opinions and so on. From the Natural Language Processing perspective, these technologies are a rich mine for tasks such as Sentiment Analysis, Automatic Humour Recognition or Opinion Mining. In this preliminary study we considered the importance of affectiveness information for blogs analysis. We focused on analysing a set of posts retrieved from a corpus of blogs related to humour in order to find the most salient items to represent the affectiveness in such sources. The results indicate that, in terms of a simple counting, there are affective features more linked to some moods than others.

1 Introduction

The importance of analysing factors related to cognitive characteristics and social phenomena through language is growing every day. Opinion Mining [8], Sentiment Analysis [18] or Computational Humour [13], are just a sample about the impact of this interest on NLP tasks. The Web 2.0 technologies have contributed to enhance the interest due to the rich mine of information that sources such as forums, blogs, wikis or social networks, represent.

In this framework, this paper is focused on blogs analysis in terms of affective features. In particular, we concentrated on analysing the posts, many of them opinions, extracted from a corpus integrated by 17,500 blogs related to humour in order to find out how affective information may characterise the emotions, sentiments, attitudes, moods, etc., profiled in such texts. Likewise, we evaluated how affectiveness impacts on blogs, opinions and reviews, in terms of measuring the presence of elements that, according to WordNet-Affect [22], imply affective information.

The paper is organised as follows. In Section 2 we underline the initial assumptions and the objective. In Section 3 we describe all the experiments carried

* The TEXT-ENTERPRISE 2.0 (TIN2009-13391-C04-03) research project has partially funded this work. The National Council for Science and Technology (CONACYT - Mexico) has funded the research work of Antonio Reyes.
out. In Section 4 we present the evaluation. Finally, in Section 5 we draw some conclusions and address the future work.

2 Humour and Affectiveness

Humour is an important instrument in our lives. It plays a significant role in physical aspects as well as sociological or psychological [11] ones. It functions as a mechanism to release our emotions; a kind of catalyst which impacts on a broad spectrum of properties linked to cognitive and social information, providing knowledge about the human behaviour\(^1\). Humour is, therefore, a human characteristic defined by the presence of amusing effects, such as laugh or well-being sensations, which are provoked by social, psychological, cognitive and linguistic factors. That is why its study implies multiple viewpoints. For instance, from a linguistic viewpoint, Attardo [1] considers humour as a phenomenon that supposes knowledge resources such as language, narrative strategies, target, situation, logical mechanism and opposition, for producing an amusing effect. By other hand, considering the research works from computational humour, it has been stated that the presence of antonyms, sexual information or adult slang are recurrent humour properties [14], also there is a trend to focus on negative elements [12], and to employ semantic ambiguity as trigger of humorous effects [19]. Likewise, taking into consideration its social properties, there are more factors that influence the perception and effectiveness of humour. For instance, affective traits: emotions, sentiments, attitudes, moods, etc.

The importance of this kind of knowledge has been noted in tasks such as opinion mining or sentiment analysis\(^2\) [6, 9, 10]. That is why we aim at studying how affective information (i.e., information regarding feelings, emotions, attitudes, etc.) may define the bloggers expression manner in terms of words that denote affectiveness in the posts, comments, and opinions extracted from blogs related to humour. We selected this kind of sources due to the following reasons:

i. despite humour seems to be a not very important topic for research purposes, it represents a recurrent and constant subject on the Internet searches, as can be noted in Figure 1. Thus, its analysis constitutes valuable knowledge to consider for different purposes;

ii. from a computational perspective, humour is an excellent mean for analysing different aspects related to NLP interests (ambiguity, irony, emotions, etc.) through different angles;

\(^1\) For instance, [20] has shown how humour appreciation is associated to personality and how this fact impacts directly on the kinds of necessary stimuli to produce humour.

\(^2\) According to the terminology discussed in [17], we differentiate these correlated areas in the fact that opinion mining suggests an “emphasis on extracting and analyzing judgments on various aspects of given items”, whereas sentiment analysis focuses on “the specific application of classifying reviews as to their polarity (either positive or negative)”. 
iii. the automatic humour processing is a task that has shown that humour may be handled by means of computers [5, 14], although it is necessary to analyse several kinds of humour;

iv. blogs are heterogeneous sites where humour is expressed in various modes: jokes, gags, punning riddles, discussions, comments and so on. This fact enhances the analysis scope to different types of humour;

v. considering the previous statements about the humour characteristics, we think that it is possible to find useful information to explain phenomena such as irony or sarcasm in order to take it into account for tasks such as opinion mining or sentiment analysis.

Fig. 1. Frequency of Internet searches related to 5 different topics during one year (August 2008 - 2009) around the world. Statistics retrieved from Google Insights.

3 Experiments

3.1 Data Sets

The experiments were carried out over a corpus of 17,500 blogs automatically retrieved from LiveJournal. All the blogs were collected taking into consideration the predefined LiveJournal mood tags [4] and, especially, the users tags: humour and joke3. This corpus was collected for automatic humour recognition purposes4. It is organised in 7 sets of 2,500 blogs according to the following categories: angry, happy, humour, sad, scared, others and general5. For our intentions of extracting the posts, we only selected the first 6 sets.

3 In the corpus construction were considered keywords such as punch line, humour, funny, laughter, laugh line, gag, joke, gag line, tag line and so on, as requisites to retrieve the blogs.


5 The sets angry, happy, sad, scared and others were retrieved employing the mood tags. The last one contains mood tags that belong to more heterogeneous categories. The set humour was retrieved employing the user tags. All of them are related to humour through keywords (Footnote 3). The set general was retrieved without considering any tag, just keywords such as news, politics, fashion, religion or technology.
Moreover, in the Evaluation Section, 5,000 opinions extracted from the TripAdvisor data set [2] and 10,662 snippets extracted from Rotten Tomatoes reviews [16] were considered.

3.2 Posts Extraction

In order to retrieve the posts from the 6 sets, we focused on identifying paragraphs with verbs related to opinions. We decided to center on verbs linked to opinion expression acts due to our further interests in analysing how humour is used to give opinions. Two requirements were established for considering a verb as an opinion verb:

i. to be linked to the \textit{opine} synonyms according to WordNet [15];
ii. to be linked to the \textit{opinion} frame according to FrameNet [3].

The sets of verbs whose lexical characteristics matched those requisites were: expect, feel, figure, guess, imagine, opine, reckon, sound off, speak up, speak out, suppose and think. The paragraphs that contained any of these verbs were extracted from the blogs and organised according to the 6 sets previously pointed out. The statistics about the paragraphs retrieved per set appear in Table 1.

<table>
<thead>
<tr>
<th>Angry</th>
<th>Happy</th>
<th>Humour</th>
<th>Sad</th>
<th>Scared</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>15,227</td>
<td>12,985</td>
<td>21,415</td>
<td>13,787</td>
<td>18,906</td>
<td>11,612</td>
</tr>
</tbody>
</table>

Table 1. Paragraphs retrieved per set according to the opinion verbs.

Given the differences in the number of paragraphs retrieved, we randomly selected 10,000 ones per set in order to keep an equivalence and to avoid any tendency based just on these differences. A total of 60,000 paragraphs containing posts, comments and opinions were considered for the further experiments.

3.3 Polarity

The first experiment was focused on identifying the positive or negative polarity in order to determine, from a sentiment analysis viewpoint, the orientation profiled in the paragraphs. We employed two public resources: SentiWordNet [7] and the Macquarie Semantic Orientation Lexicon (MSOL) [21].

The first one proposes a set of graduated tags to cover positive and negative ranges for the following morphosyntactic categories: nouns, verbs, adjectives and adverbs. We only selected nouns, verbs and adjectives, if and only if, they passed an empirically founded threshold $\geq 375$ in their positive or negative scores. For instance, the adjective \textit{adventive} has a negative score of 875 and a positive score of 0.0, whereas the adjective \textit{rash} has a positive score of 625 and a negative one of 0.25. The items with scores $\leq 375$ or with equal scores were not taken into consideration because they do not provide sufficient information to define an orientation. The MSOL contains 76,400 entries (30,458 positive and 45,942...
negative ones) and considers a Roget-like thesaurus for expanding their positive and negative seeds. According to the authors [21], it reaches a higher-coverage regarding the phrase polarity than SentiWordNet.

We determined a global polarity per set on the basis of the presence of positive and negative items registered in SentiWordNet and MSOL. The process consisted in the following steps:

i. let $S_p$ be the set of positive items registered in any of both resources;
ii. let $S_n$ be the set of negative items registered in any of both resources;
iii. let $p_1, \ldots, p_x$ be every one of the posts in category $C_k$ containing any opinion verb;
iv. if $\sum S_p > \sum S_n$ given $p_x$, then the post was labelled with positive polarity;
v. if $\sum S_p < \sum S_n$ given $p_x$, then the post was labelled with negative polarity;
vi. otherwise, $p_x$ was labelled as neutral polarity.

According to this process, we found out a substantial trend to focus positive orientation on 80% of the paragraphs, regardless of the resource or the set they belong. This can be viewed in Table 2.

Furthermore, all the positive and negative items identified in the paragraphs were retrieved in order to detect the most recurrent words. These elements were tagged according to their morphosyntactic category. We just concentrated on the 100 most frequent elements belonging to the noun, verb and adjective categories for determining what elements appeared most frequently. Surprisingly, despite the strong positive trend, among the 100 most recurrent items appear more words belonging to the set of the negative ones, being the negative adjectives the elements which appear quite often (for instance, in the angry and humour sets, 46 of 100 words are negative adjectives). However, considering the results in Table 2, it is clear that this behaviour only impacts on these 100 most frequent words.

Finally, on the basis of the results obtained, we may suppose that: a) considering the strong positive polarity, and beyond the 100 most frequent words, the negative items appear fewer times than the positive ones alongside the paragraphs; b) the negative items are concise referents, that is why they appear very
few times. It is sufficient one concise negative word for getting a negative mean-
ing (for instance, an insult); c) if the positive items appear quite often in the
texts, perhaps it is due to the whole meaning is meant to produce the opposite
effect, that is, we are considering the presence of irony or sarcasm\(^6\).

3.4 Affectiveness

In this experiment we focused on finding out what kinds of emotions are the most
profiled by the bloggers in their posts. The experiment consisted in computing
the amount of affective elements, based on the WordNet-Affect categories\(^7\), per
every one of the 60,000 paragraphs.

![Categories coverage for the posts labelled with positive polarity.](image)

The analysis was performed considering separately positive, negative and
neutral paragraphs. That means we aim at finding the most important emotion
or sentiment profiled in the posts regardless their polarity\(^8\). Moreover, given the
inferences noted in the previous section, we want to study further, irony and sar-
casm. The process consisted in computing, for every single post belonging to the
positive, negative and neutral polarity, all the affective categories it contained.
Later, we weighted the occurrences according to the 6 categories. The results
obtained are not very useful due to the unbalanced representativeness of some
classes. For instance, the items of the categories \textit{eds} or \textit{phy} are much fewer than
those ones of the categories \textit{emo} or \textit{tra}. This can be noted in Figure 2, where we
represented the classes coverage, in terms of occurrences, for all the paragraphs
labelled with positive polarity.

\(^6\) In Appendix A we provide six examples taken from the posts labelled with positive
polarity which may be analysed as ironic ones.

\(^7\) The affective information represented by these categories is attitude (att), behaviour
(beh), cognitive state (cog), edonic signal (eds), emotion (emo), mood (moo), physi-
cal state (phy), emotional response (res), sensation (sen), emotion-eliciting situation
(sit) and trait (tra). All the information about the concepts symbolised by these
categories appears in [22].

\(^8\) For practical reasons, the experiment was only performed over the polarity results
obtained with SentiWordNet.
Therefore, in order to avoid not informative categories, we applied an information gain filter [23]. The most informative categories, according to this measure, were: att, beh, cog, emo and tra. Considering just these classes, we repeated the experiment. The results appear in Table 3.

<table>
<thead>
<tr>
<th>Negative</th>
<th>Angry</th>
<th>Happy</th>
<th>Humour</th>
<th>Sad</th>
<th>Scared</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>att</td>
<td>78</td>
<td>64</td>
<td>122</td>
<td>59</td>
<td>95</td>
<td>58</td>
</tr>
<tr>
<td>beh</td>
<td>276</td>
<td>170</td>
<td>388</td>
<td>184</td>
<td>269</td>
<td>210</td>
</tr>
<tr>
<td>cog</td>
<td>382</td>
<td>332</td>
<td>471</td>
<td>262</td>
<td>398</td>
<td>205</td>
</tr>
<tr>
<td>emo</td>
<td>651</td>
<td>517</td>
<td>936</td>
<td>534</td>
<td>728</td>
<td>443</td>
</tr>
<tr>
<td>tra</td>
<td>889</td>
<td>690</td>
<td>1080</td>
<td>642</td>
<td>883</td>
<td>582</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Positive</th>
<th>Angry</th>
<th>Happy</th>
<th>Humour</th>
<th>Sad</th>
<th>Scared</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>att</td>
<td>757</td>
<td>832</td>
<td>1,331</td>
<td>704</td>
<td>1,112</td>
<td>583</td>
</tr>
<tr>
<td>beh</td>
<td>2,764</td>
<td>2,589</td>
<td>4,627</td>
<td>2,748</td>
<td>4,196</td>
<td>2,121</td>
</tr>
<tr>
<td>cog</td>
<td>3,862</td>
<td>3,202</td>
<td>5,642</td>
<td>3,687</td>
<td>5,716</td>
<td>2,862</td>
</tr>
<tr>
<td>emo</td>
<td>7,727</td>
<td>6,688</td>
<td>12,496</td>
<td>8,139</td>
<td>12,201</td>
<td>5,833</td>
</tr>
<tr>
<td>tra</td>
<td>15,031</td>
<td>13,357</td>
<td>17,322</td>
<td>14,596</td>
<td>16,951</td>
<td>11,698</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Neutral</th>
<th>Angry</th>
<th>Happy</th>
<th>Humour</th>
<th>Sad</th>
<th>Scared</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>att</td>
<td>41</td>
<td>23</td>
<td>54</td>
<td>27</td>
<td>49</td>
<td>19</td>
</tr>
<tr>
<td>beh</td>
<td>140</td>
<td>126</td>
<td>181</td>
<td>111</td>
<td>123</td>
<td>100</td>
</tr>
<tr>
<td>cog</td>
<td>178</td>
<td>171</td>
<td>202</td>
<td>137</td>
<td>183</td>
<td>131</td>
</tr>
<tr>
<td>emo</td>
<td>262</td>
<td>256</td>
<td>451</td>
<td>244</td>
<td>398</td>
<td>230</td>
</tr>
<tr>
<td>tra</td>
<td>403</td>
<td>395</td>
<td>674</td>
<td>380</td>
<td>515</td>
<td>315</td>
</tr>
</tbody>
</table>

Table 3. Affectiveness per set and polarity.

According to the results obtained with the most informative categories, we can notice that the most profiled features by the bloggers tend to express emotions and, especially, affective traits. This can be corroborated with the information presented in the table. Regardless the polarity, both emo and tra categories are the items which occur most frequently in the 6 sets. Moreover, it can also be appreciated that humour and scared categories are the sets which concentrate more elements regardless the affective category.

4 Evaluation

In this section we aim at identifying the most profiled affective elements beyond the posts. On this basis, we selected 5,000 opinions from the TripAdvisor corpus and 10,662 snippets (one half labelled with positive polarity and the other one labelled with negative polarity) from the Rotten Tomatoes reviews. All these data were processed following the affectiveness experiment for determining the amount of affective items.

The underlying assumption was to determine whether or not the affective categories appear similarly in discourses related to travel and movies than in discourses related to humour. The results are shown in Table 4.
Table 4. Assessment per data set

According to the simple presence of affective elements reported in this table, it is evident that the most profiled affective category is the one linked to affective traits, as in the previous experiment. However, there is a difference in the following profiled categories: the emotion and behaviour ones alternate, in terms of occurrences of their elements, both in opinions and snippets. A distinct situation than the one of the previous experiment, in which it is obvious that the profiled categories are only the trait and emotion ones.

5 Conclusions and Future Work

In this preliminary study we have tried to identify what are the most salient affective categories profiled in 60,000 posts extracted from blogs related to humour. Likewise, we compared whether these categories appeared steadily in opinions and reviews. The results suggest that, despite the amount of data per set, its polarity or the topic treated, there is a constant or trend in expressing affective traits. Likewise, it seems that the emotion category plays an important role to communicate affective information in the discourses related to humour.

Furthermore, it is important to note that, regarding the polarity experiment, there is an important trend to focus positive elements in the 6 initial categories (angry, happy, humour, sad, scared, others). This fact will be studied further in order to analyse the assumption c), mentioned in Section 3.3: irony and sarcasm. Finally, by means of considering the thematic roles of the opinion verbs, we also plan to study whether or not it is possible an automatic topic identification.

References

Appendix A: Ironic posts?

1. Now it’s true that I have a tendency to dislike most of the things related to the 60s. I can’t help it. But I still tried to understand what was so great about this book, there was no point, I hated this book from the very beginning to the end. Pynchon’s story is probably full of irony but I couldn’t possibly feel it when I was too busy trying to concentrate on this awful writing.

2. The ongoing ridiculous situation brewing between bloggers and the Associated Press has now taken a turn towards the enjoyably hilarious. We had already mentioned the fact that, despite the AP’s complaints that bloggers quoting less than 100 words were violating fair use, the AP had a long history of quoting more than 100 words from bloggers – and not even linking back to the original blog. Now, the AP’s own article about this brouhaha quoted (without linking) twenty-two words from TechCrunch. That’s 18 words more than the supposed four word “limit” the AP has suggested. With an ironic chance that wide, TechCrunch’s Michael Arrington couldn’t resist, and asked his lawyer to send a DMCA take down notice to the Associated Press, along with a bill for $12.50 (directly off the AP’s own pricing schedule). He admits that it’s ridiculous, but that’s what his actions are designed to present. By law, the AP should be required to take down the content before filing a response – though, since it’s filing the response to itself, then perhaps it won’t need to take down the content. Either way, this helps illustrate the insanity of the entire situation.

3. Okay, sports fans, it’s official – Cleveland weather sucks ass! Opening day at Jacobs Field (where all of AG.com was headed tomorrow as a company reward/party) has been cancelled, since we are now expecting 3-6 inches of snow! They are rescheduling the game for Tuesday afternoon, so we will probably at least get to go... Though in a sharp twist of irony, the conference call that I asked people to reschedule is going to have to be rescheduled again (originally on Monday due to forgetfulness, it was moved to Tuesday afternoon – ) d’oh?! Ha ha. Okay, maybe it’s only funny to me, and a handful of people at work who (probably) don’t read this.

4. And it’s a darn good thing my Monday ended on such a good note, because Tuesday sure as hell started as a train wreck. I really need to get better at detecting when people are bullshitting me. I really hate to just assume that everyone is bullshitting me, because that level of cynicism doesn’t really jive with my core Pollyanna-like personality. On the other hand, I think that tending to believe people might not be the best thing either.

5. My landlady’s still surreal and psychotic but that’s no real news. I think she’s pissed that ALL four of the cats now simply hang out in my room all the time. All of them sleep on my bed (or, in Chloe’s case, under it) and hang out in the bathroom while I’m chilling in the tub. She still thinks I read most nights and listen to music instead of watching TV in a dark room like she is. She will get up at 8am on Saturday and spend ALL DAY IN FRONT OF THE TV. I think I just fear BECOMING LIKE HER. Giving up. Cocooning. I really hate to just assume that everyone is bullshitting me, because that level of cynicism doesn’t really jive with my core Pollyanna-like personality. On the other hand, I think that tending to believe people might not be the best thing either.

6. I’d like to abolish the insidious terms Darwinism, Darwinist and Darwinian. They suggest a false narrowness to the field of modern evolutionary biology, as though it was the brainchild of a single person 150 years ago, rather than a vast, complex and evolving subject to which many other great figures have contributed. (The science would be in a sorry state if one man 150 years ago had, in fact, discovered everything there was to say.) Obsessively focusing on Darwin, perpetually asking whether he was right about this or that, implies that the discovery of something he didn’t think of or know about somehow undermines or threatens the whole enterprise of evolutionary biology today.