Opinion Mining on YouTube - tech review

Introduction

Social media text analysis is an important study because media such as Youtube and Twitter affect millions of users. Youtube comment analysis, however, is especially more difficult due to the fact that there are comments can either be about the video itself or the product the video is about, and there are also irrelevant comments that shouldn't be included in text analysis. This paper introduces a systematic approach on Opinion Mining (OM) that shows significant improvement compared to pre-existing approaches regarding polarity and type classification of comments.

Body

In this paper, we learn about the algorithm and training process of the model in order to classify comments by 1) whether they are about the video or the product that the video is about, and 2) whether the sentiment of the comment is negative, positive, or neutral. The previous algorithms developed that are mentioned in the paper are not very effective because the nature of comments in social media is that they are short and have many inconsistencies and errors, whether deliberate or accidental. The new approach uses syntactic structures to adapt to these types of comments, and also utilizes tags in order to generalize across different domains.

- Representations/models

Not all comments on a video are relevant to the content, which means that they should be filtered out so that they don't contribute to the count of positive or negative sentiments towards the video. After filtering out such comments, in order to initialize some sentiment words, there are multiple "bag of words" being used.

- Structural model

This OM algorithm uses a simple shallow 2-level tree structure built from word lemmas and part-of-speech tagging, which are eventually grouped into chunks. The topic, or "concept", of the video is retrieved from the video title and description, which are used to match against the tree to determine the relatedness of the comment to the video.

The example used in this paper is the sentence “iPad 2 is better. The superior apps just destroy the xoom” for a video describing the product xoom. When we think about a very simple algorithm that counts negative and positive words, we would have 2 positives (“better” and “superior”) and 1 negative (“destroy”), which would lead the algorithm to think that this particular comment is a positive one. However, this is not the case - the comment is actually expressing negative sentiments against xoom. The OM algorithm instead analyzes this comment in a way that associates “destroy” with xoom, and “better” with iPad 2. This is done by using the STRUCT model that encodes each sentence into a tree. Since the video is about xoom, the algorithm in this case will mark the comment as a negative one since it praises iPad 2 and denounces xoom. Also, since the video title and description include “xoom”, this comment is marked as a comment for the PRODUCT, not the video itself.

- Experiment

With this simple algorithm, testing was done on two product categories: AUTO and TABLETS. Videos were split 50/50 between training and test sets, with all comments for a single video being in a single set (i.e. comments were not split between the sets). There were approximately 35k comments total. The results show that “video-positive/negative” were the most difficult to predict because they were the most scarce classes. For the training accuracy, for both AUTO and TABLETS, the training accuracy increased from 1k to 4k (increments of 1k), although there wasn’t a significant increase from 4k to all comments.

Conclusion

With a dataset that has ironic or irrelevant comments filtered out, this OM algorithm works pretty well because it utilizes the video topic and part-of-speech tagging in a way that the sentiment of the comment can be attached to either the video itself or the product. However, this can mainly only be applied to videos that have the appropriate video title and description and with comments that don’t contain irony or grammar irregularity. The use of bag-of-words is also a big unstable dependency especially when it comes to comments that mean the opposite of what it states. As stated in the Future Work section, perhaps multi-label classifiers will increase the algorithm accuracy for all video comments, and (maybe with some machine learning) we can incorporate “human knowledge” to detect the irony in a comment, although this opens up a whole new debate as to what piece of knowledge is “correct” and the algorithm should learn.

Reference

- https://research.google/pubs/pub42471/