International Journal of Sciences

Research Article Volume 4 – June 2015 (06)

Creating Sentiment Corpora from Microblogging using Machine Learning

A. Alghamdi[1]

[1]Albaha University, Saudi Arabia

Abstract: This paper describes our work-in-progress implementation of a system to create microblogging sentiment corpora. We developed a smart phone application that aims to create microblogging lexical resources which can be used for sentient analysis. Our proposed approach to machine learning is to use active learning which aims to make the process of resource creation more effort and time efficient. Active learning helps us identify instants that if labeled will improve our sentiment analyser. Microblogging, on the other hand, reduces the cost of creating the resources.

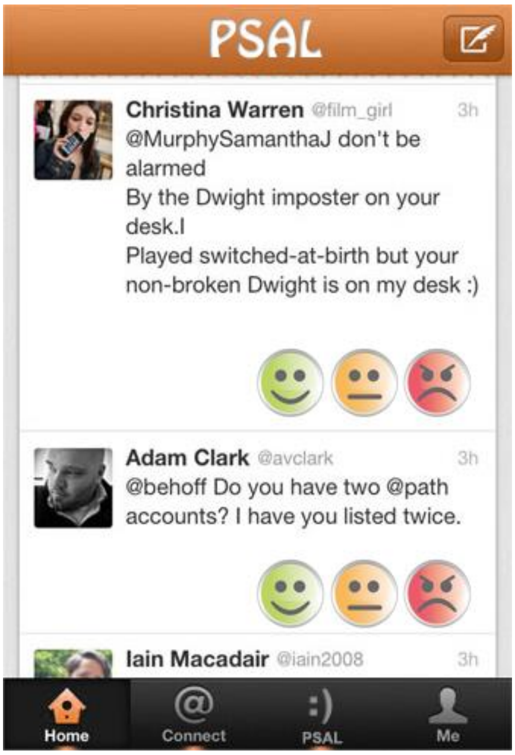
Introduction

Natural Language Processing (NLP) techniques such as sentiment analysis have developed at a dramatic pace in recent years, partly attributable to the fact that many annotated resources have become large in scale. These resources are created by humans and require intensive effort. Although some lexical resources for sentiment analysis might be sufficiently avail- able, however, more specific resources such as microblogging lexical resources for sentiment analysis are still needed. Human annotation for such corpora are always difficult, costly, and resource-hungry processes.

With the emerge of crowdsourcing as a successful approaches to create resource, creating more specific annotated resources and getting people to contribute to the creation, annotation or validation of these resources became feasible.

As Active Learning (AL) (Settles, 2009) joins the scene, drastically reducing the amount of annotation required to train a highly accurate sentiment analyzer becomes feasible. Even though active learning looks promising being discussed in many research papers, it still struggles to find its way to real world applications (Attenberg and Provost, 2011). Moreover, most applied research on Active Learning involving crowdsourcing does not really pay attention to the nature of participants, the social elements, or life as an interactive environment.

Personalized content made extraordinary improvs how in people interact with resources. Social networks receive most of that interaction and make people more connected, committed, and even addicted sometime to technology. Therefore, we believe that active learning and crowdsourcing should also follow the same methods to become personalized and social. We assume that if users receive more personalized annotating requests, they are more likely to con- tribute and produce more accurate annotations. Moreover, the social atmosphere is more attrac- tive for users and can encourages them to partic- ipate in the loop. Previous attempts shows that personalization of active learning is essential to obtain better results (Harpale and Yang, 2008).



Christina Warren @film_girl 3h
@MurphySamanthaJ don't be alarmed
By the Dwight imposter on your desk.I
Played switched-at-birth but your non-broken Dwight is on my desk :)

Adam Clark @avclark 3h
@behoff Do you have two @path accounts? I have you listed twice.

Iain Macadair @iain2008 3h

Figure 1: PSAL application interface

Published at: http://www.ijsciences.com/pub/issue/2015-06/
Article Number: V4201506760; Online ISSN: 2305-3925; Print ISSN: 2410-4477

A. Alghamdi (Correspondence)
ansdaif@gmail.com
☎ +

We propose Personalized and Social Active Learning (PSAL), a framework to create *microblogging lexical resources* for sentiment analysis with crowdsourcing and active learn- ing while being personalized and social. Active learning will help us identify entities that, if labeled, will improve the performance of the sentiment analyzer. On the other hand, crowdsourcing will reduce the cost of resource creation. Moreover, personalized content and the social environment will help make creating re- sources a fun and enjoyable task. Our work makes the following contributions.

- We propose a smartphone microblogging personalized and social application called "PSAL: People who Smile A Lot," to tie active learning and crowdsourcing in a personalized and social framework.
- We propose a new paradigm to creating high quality and low cost resources, to be used to train sentiment analyzers. The remainder of this work-in-progress paper is be organized as follows. Section 2 provide an overview of how our project works, primarily the enjoyable aspects including personal and social elements to be expressed. The back-end technical part explaining our NLP/IR processes are addressed in sections 3. Lastly, concluding in section 4.

Game With Purpose

Crowdsourcing implementations alway make the best attempts to keep users engaged. A number of different approaches attempt to catch people's attention and motivate them to participate and vary from paying users (e.g., MTurk) to keeping them entertained (e.g., game with a purpose (Von Ahn and Dabbish, 2004)). It is essential for projects such as ours to have their own flavor of entertainment and enjoyment; therefore, our work will be built from scratch with that in mind.

Our project will be built as a layer on top of Twitter. We are not aiming to reinvent the wheel and create a new social network. We will attempt to mimic the Twitter official smart-phone application experience, however, our application will have its own added flavor.

Users are able to log in with their account and view their own timeline. As shown in Figure 1, next to every tweet are three emotion faces. These smily faces allow users to choose whether the tweet is positive, normal, or negative. Users are able to click only one of these emotion faces. They also can ignore any tweet they wish.

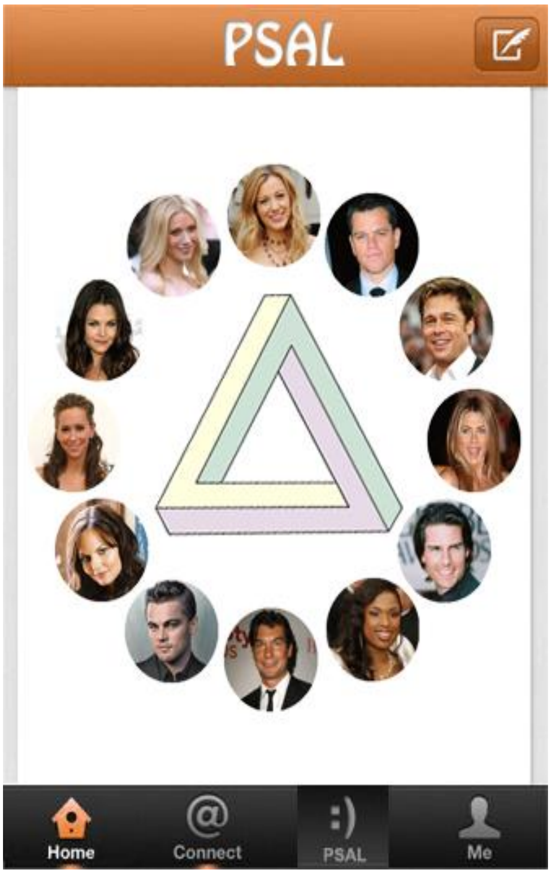


Figure 2: PSAL view, user can click on any of the profile photos to tweet PSAL tags

By the end of a specific period, users can see who in there timeline are "people who smile a lot" (Figure 2). A triangle with three colors denoting the three sentimental tags: positive, normal, and negative. However, our proposed approach here is not to show tweets themselves, but to show the people how tweeted them. An analysis curried out at the background classify- ing people who there tweets ware tagged and show them as a circle around the sentiment triangle. Creating what we like to call a "PSAL tag.". This brings us to the social elements of our project.

PSAL tags Because Twitter already provide many social elements, we use them allowing users to share their PSAL tags with their followers. Users will be able to tweet their PSAL tags that *mention* "people who smile a lot" ac- counts. We hope that will encourage more users to take part and PSAL tag their timeline, particularly if they have been *mentioned* by others. By that time, our project will start to learn from users' behavior. Suggesting a sentimental tag for every tweet and suggesting who and what to tag, which introduces the personalization experience of our project.

Personalization We do not expect users to tag every single tweet on their timeline; other- wise, doing so will be boring and a waste of time. Our project is to have its own personal feel in the sense that it learns users preferences and adjusts itself according to that. Users will be given some suggested sentimental tags based on who they normally interact with, what topics they are inserted at, and other PSAL tags created by other users. Creating such personalized feedback request will improve the engagement and interaction between users and our active learner.

Experiment Implementation

Our back-end will be recording and creating resources. Different components will create our pipeline, from sentiment analyzer to personalization machine learning algorithms are to run in the background to create an experience that our users will enjoy. Our back-end primarily consists of two modules:

Crowdsourcing Framework: The power of crowdsourcing has been dominating the Web. Examples such as WikiPedia and others are the result of such an approach. Our crowdsourcing approach varies in the fact that we do not rely only on uses, and we are attempting to make the user experience more enjoyable. As users start to use our application, they are creating microblogging lexical resources for sentiment analysis. These resources will also be used for our sentiment analysis active learner to help users tagging tweets. Because our crowd- sourcing environment is build on top of Twitter, it is by default a kind of social network. However, as we are not asking users to tag every single tweet, to minimize the needed effort, we are relying on active learning to select only the most valuable tweets and ask for feedback. User interaction such as PSAL tags, mentions, hashtags, and even feedback requests that are ignored are recorded to provide more personalized experience. All data are to be recorded anonymously, and users permeation will be taken in advance.

Active Learning: Our second important component is the personal and social active learning approach we are proposing. Active an- notation (Vlachos, 2006) (the application of active learning) will help to dramatically reduce the amount of necessary annotation by asking users to tag only the most useful data for our classifier. Active learning have been always static and not really *active* in the sense that it is supposed to interact continuously with annotators. We believe active learning should have some sort of relationship with taggers that form there interaction. It is the active learner who are supposed to adjust itself and work as annotators expect. Therefore, our proposed personalized and social active learning is to bridge that gap.

Our seed examples to train our classifier are created by uses, and are to be extended every time users tag more data. Right new we are thinking of retraining our tagger overnight, however, we may conceder more appropriate methods. The next step is to apply our trained classifier to more unlabeled data, tagging more resources and storing them. Next, our pipeline is to determine additional examples that the classifier views as more informative but not certain how to classify. A number of approaches have been proposed for sample selection, and the one we will be using is *uncertainty sampling* (Lewis, 1995), which is the most commonly used query framework (Settles, 2009).

The selected uncertain samples are filtered and personalized, creating more personal feedback requests to users. After that data are tagged, the cycle is repeated, allowing the learner to quickly refine the decision boundary between the classes.

Conclusion

The study is a work-in-progress pilot experimental attempt to define our proposed paradigm, personal and social active learning, aiming to improve active learning experience and quality. Our work is still in its very early stages. However, we are to release a real-world implementation and we believe feedback at this particular stage will help us minimize the needed iterations toward having more mature implementation of our idea.

References

1) Josh Attenberg and Foster Provost. Inactive learning?: difficulties employing active learning in practice. *SIGKDD Explor. Newsl.*, 12(2):36–41, March 2011. ISSN 1931-0145. http://dx.doi.org/10.1145/1964897.1964906
2) A.S. Harpale and Y. Yang. Personalized active learning for collaborative filtering. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in in- formation retrieval*, pages 91–98. ACM, 2008. http://dx.doi.org/10.1145/1390334.1390352
3) David D. Lewis. A sequential algorithm for train- ing text classifiers: corrigendum and additional data. *SIGIR Forum*, 29(2):13–19, September 1995. ISSN 0163-5840. http://dx.doi.org/10.1145/219587.219592
4) Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, Uni- versity of Wisconsin–Madison, 2009.
5) Vlachos. Workshop on adaptive text extraction and mining. In *EACL 2006*, Trento, 2006.
6) L. Von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM, 2004. http://dx.doi.org/10.1145/985692.985733