User-Supplied Sentiments in Tweets

Samuel Huron, Jean-Daniel Fekete

User-Supplied Sentiments in Tweets

Samuel Huron
INRIA / IRI
samuel.huron@cybunk.com

Jean-Daniel Fekete
INRIA
Jean-Daniel.Fekete@inria.fr

ABSTRACT

Microblogging has become very popular among web users: Twitter broadcasted 200 millions of them every day in 2011. These tweets can be used for social studies, opinion mining, and sentiment analysis. Tweets have been effectively used to analyze opinions and sentiments from specific events such as TV political debates or conference presentations. To analyze tweets, researchers either use Natural Language Processing or human-analyses, *e. g.* with Amazon Mechanical Turks. In this paper we describe a third method that can replace or complement the first two: convincing the tweet authors to explicitly express their opinion using a simple unambiguous syntax—a *tag*—while they tweet.

We report on our experience with the PolemicTweet system where authors tag the tweets they send during conferences and TV shows. We explain how we implemented PolemicTweet and discuss briefly the pro and cons.

We believe that author-tagging technologies can effectively complement automated analysis, in particular for analyzing tweets that are by essence ambiguous, and could be ironic, sarcastic or cryptic. We describe the incentive we used to convince the authors and the result we obtained.

Author Keywords

live event, Twitter, tagging, video annotation, crowdsourcing, Natural Language Processing (NLP).

ACM Classification Keywords

H.5.3 Information Interfaces and Presentation: Group and Organization Interfaces

General Terms

Human Factors; Design; Measurement.

INTRODUCTION

In the recent years, microposts such as *tweets* have been used to analyze data about particular events, such as political opinions during political TV debates [6], sport events [19], and crisis coordination [4]. Analyzing tweets is not only useful for the sake of understanding what people think about particular topics in general, it also informs about particular events happening at some place in the world (*e. g.* a tornado hitting a place) or broadcasted on TV and visible on a large scale (*e. g.* the Olympics or a political TV debate). As such, tweets—as well as other social networks such as Facebook—provide real-time sources of annotation for spatio-temporal events.

To analyze tweets, researchers have been using Natural Language Processing (NLP) or human-made analyses *e. g.* with Amazon Mechanical Turks. While useful, NLP analyses suffer from several pitfalls:

- Their cost is proportional to the number of messages to process and can become substantial;
- Most NLP research is targeted at English and not usable with other languages that still use tweets heavily and are of great interest (*e. g.* Arabic speaking countries);
- Even for English, sentiment analysis or topic classification on tweets is a complex process without "off the shelf' solution yet;
- NLP methods suffer from precision[1] problems due to the frequent use of irony, sarcasm, slang, and abbreviations.

Human analysts are better at understanding the language than NLP but are more costly both in time and price. They are not immune of precision problems due to the ambiguity of the natural language and the difficulty of interpreting irony, sarcasm, slang, and abbreviations.

In this position paper we describe a third method that complements the first two: persuading tweet authors to express their intents using an unambiguous coding—a *tag*—in their microposts. This method avoids having to interpret the intents of the tweets afterwards. In our current system, we ask authors to express 4 intents: agreement, disagreement, a question, or that they provide a reference information (*e. g.* a quote) related to a topic. To be effective, our method requires a careful mix of incentive and simplicity, as reminded by Shipman et al. [17]: without incentive, people will not spend more time adding explicit statements in their microposts; if expressing their intents is complex or cumbersome, they will not do it either.

Our method enriches the micropost's stream with information that is cheap, easy to parse, work in every language, and is not sensitive to bad interpretation. We report on our experience with author-tagging of conferences and TV shows, we explain how we implemented our method, and discuss its

[1] We refer to precision and recall as in Pattern Recognition or Machine Learning. *Precision* is the fraction of retrieved instances that are relevant (*i. e.* well classified), while *recall* is the fraction of relevant instances that are retrieved [14].

pro and cons. The method has been used for about 30 conferences so far and a national TV shows in the last two years. Our method—because it is cheap and reliable—can be deployed to provide an initial set of annotations related to space and time for visual analytics.

RELATED WORK

NLP for Sentiment Analysis on Tweets

In a few years, NLP techniques on tweets have improved greatly due to the availability of large corpora and the design of appropriate language models with short words and non standard English [13]. O'Connor [12] shows that lexicon-based classification techniques have a very low recall when using a lexicon of standard English while many tweets are written in an informal dialect. Today the combination of different techniques, such as machine learning based approach [1], semantic rule based approach [11], and graph based optimization [18], provide a significant improvement of sentiment classification, achieving about 86% of accuracy [11].

Still, NLP suffers from the pitfalls described in the introduction: a cost proportional to the number of messages to process, limited to English, not readily available from "off the shelf" product, and limited in recall.

Human Annotation via Crowdsourcing

According to MITNews [8] "Crowdsourcing is a technique for processing a task over the internet by splitting it into small chunks that dozens, hundreds or even thousands of people complete." This technique is particularly useful for tasks that are trivial for humans but difficult, for computers [15]. Crowdsourcing systems have demonstrated their effectiveness with several applications such as photo selection [2], data analysis [20], and question answering services [3].

Diakopoulos and Shamma [7] have crowdsourced sentiment classification for tweets using Amazon Mechanical Turks. Turkers were paid $0.05 for ten classifications and have to recognize 4 types of sentiments: negative, positive, mixed, and "other". The corpus of tweet was in English, on a binary choice topic (2008 US presidential debate). In a second article [6], they applied machine-learning algorithms to perform the same analysis with lower precision and cost but higher speed. Crowdsourcing sentiment classification on tweets is now a standard service provided by the CrowdFlower [5] crowdsourcing platform, although it is limited to English.

Crowdsourcing suffers from the following pitfalls: *cost, speed,* (it is much slower than automated algorithms), and it is not immune from *ambiguity*. We have designed PolemicTweet [9] to address some of these pitfalls.

PolemicTweet

PolemicTweet (PT) was developed in response to the need for tagging videos of conferences to facilitate their browsing and analysis. A large number of organizations (universities, companies, institutions) organize middle-sized events, from workshop to conferences, gathering 20 to 300 attendees, each session lasting about 1 to 2 hours.

Many of these events are video recorded for archival or for sharing to a broader audience. When we started to develop PolemicTweet, automatic sentiment classification methods on tweets were not as effective as they are now, at least for English; since PT is mainly deployed in France, automatic sentiment classification is still not an effective option.

Therefore, PolemicTweet relies on tweet authors to insert special tags in their tweets to express their intents with regard to the presentation. PolemicTweet currently supports 4 intents, expressed by inserting two repeated characters anywhere in the tweet: agreement `++`, disagreement `--`, question `??`, and reference material added `==`. This information is collected and used to annotate the produced video (Figure 1), as can be seen on the web site [16] that already contains about 30 annotated videos.

The colored bar chart added to the video navigation panel is a navigation aid that shows every tweet about the event as a colored 5 pixels square. The color code is straightforward: green for agreement, red for disagreement, orange for a question, and blue for reference material added. Grey squares represent tweets with no PolemicTweet tag.

Protocol

To increase the chance that participants use our syntax in their tweets, we have designed a special social protocol and some tools. It is difficult to know precisely how much each of these contribute to convincing the participants but, overall, we obtain an average around 40% and a peak at 77% when the audience is very active. We consider the protocol crucial to the success of PolemicTweet: incentives must be used to convince tweet authors to use the syntax.

The social protocol consists of:

Before the event we have designed a "connection package" to make sure the audience is properly informed. It consists of a flyer, given to attendees during the registration process, containing information about the wireless network, the PT syntax, the URL of the event web page, and the Twitter hashtag for the event.

During the event the web page of the event can be used to read and send tweets with support for the syntax. It includes colored feedback about the tweets that include the syntax. We also record all the tweets that use the defined hashtag.

After the event a tweet is sent to the participants notifying them of the availability of the recorded video on the web site and thanking them for their participation. The video captured during the event is published on the PT web site with the visualization of the tweets activity.

We have implemented two tools to support our protocol and process:

A PT-enhanced tweet client: a web application that allows reading and sending tweets with support for our syntax (Figure 2). By visually highlighting tagged tweets in real time, this web application creates an awareness of the other opinions and incentive to tag her/his own tweets.

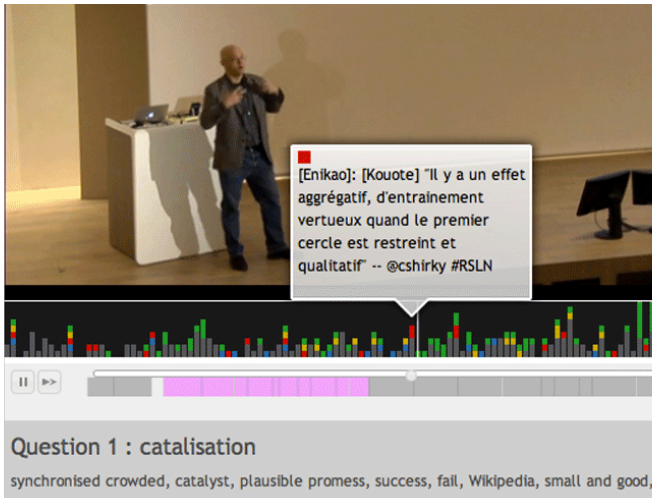

Figure 1. The Metadata Player: a Video Player with the Polemic Timeline.

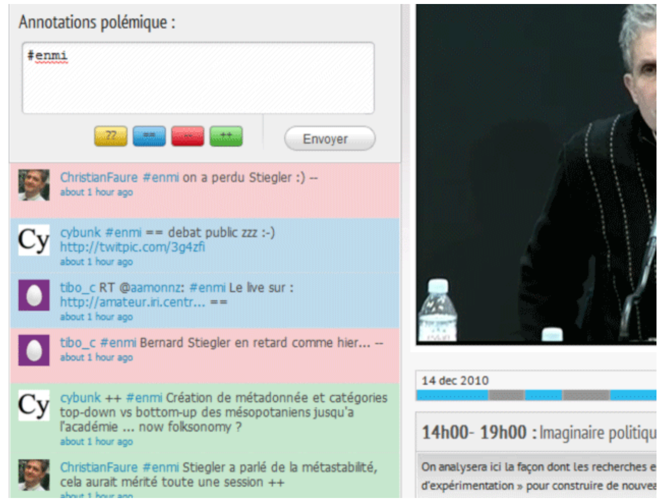

Figure 2. Enhanced Tweeter client used during the live conference to visualize tweets and remind the PT syntax.

The Metadata Player: a web application that extends existing video player by adding the color tweets bar chart (Figure 1) [**?**]. This video player allows navigation on the video using either the standard controls or by interacting with the tweets. By moving the mouse over a square, the text of the tweet is displayed on a popup window. Clicking on a square cues the video at the time it was sent.

DISCUSSION

We designed to PolemicTweet to annotate tweets but is has become a more complete system, because the live feedback provided by the tweets reader in our client adds a lot to the experience of the audience. Participants can share their feelings with others, get a quick overview of the mood and reactions, organizers can gather feedback, questions and reference material about the talks, and presenters can collect feedback on their talk too. So far, the feedback has never been insulting or embarrassing, even if, at times, comments have been negative. PolemicTweet is used on a regular basis and other institutions have been asking to use it for their own events, which is a good sign of success.

The cost of PolemicTweet is different than using NLP or human annotation: it is mostly independent of the number of tweets since tweet authors enter the annotations with no extra cost for the conference organizers. The quality is also different: less ambiguous since authors express their intents, but with only 40% of annotated tweets. As raised in the introduction, nothing prevents NLP or human annotations to be performed to the remaining tweets if needed.

PolemicTweet has been also used to annotate national TV shows, under the name *Bubble TV*. We have modified the social protocol and tools to present a more appealing interface with animations to be attractive on TV. Here again, tweet authors have used a variant syntax for the tags so the resulting stream of tweets can be used to create an annotated a time line of each show. In that particular case, the goal was more to receive live feedback about the show presenters than to annotate it. Still, both feedbacks are useful and usable with no extra cost.

With our experience deploying PolemicTweet, we see it as a practical tool to generate an annotated timeline to help the exploration and analysis of recorded events. However, it also has limitations.

Limited number of tags: tweet authors will never remember more than a few tags: PolemicTweet cannot provide richly encoded annotations. Yet, other methods based on NLP or crowdsourcing have been very limited too up to now, so PolemicTweet remains useful and could be complemented by entity recognition or other technologies when they become robust enough on tweets.

Tags should be known in advance: exploration of topics and alternative sentiments on tweets cannot be done with our technique, and changing the tag set during the event would certainly confuse the participants and lower the quality and quantity of tagged tweets.

Non technical communities are excluded: we noticed a strong variation on the number of tagged tweets when the audience changed from high-tech communities to more traditional ones.

CONCLUSION AND FUTURE WORK

This position paper explains how we implemented PolemicTweet, a system inciting people to tag their tweets with unambiguous expressions of some of their intents. PolemicTweet specifies a simple syntax for agreement, disagreement, questions, and the adding of reference material in a tweet.

The deployment of PolemicTweet for a series of about 30 events during 2 years shows the practicality of our approach and very positive results in annotating videos with a limited amount of effort. Beyond the relative success of our system, we have witnessed a great interest from our audience and from other event organizers; we assume it is due to the increased engagement offered by PolemicTweet.

There are many events and text posting systems that could benefit from user-supplied sentiments to mine opinions in real time, *e. g.* social TV, classrooms, or public debates. Our technique would provide some level of unambiguous feedback to the audience and organizers. This position paper also introduced a first step towards the real-time exploration and classification, of sentiments using visualization. We believe our method is applicable to a wide variety of different scenarios and events.

For future work, we would like to test if we can apply this type of method to more complex tasks (*e. g.* sharing classroom annotations). We also would like to better understand the degree of incentive induced by each component and how to improve it, as well as how we can enhance the degree of tag expressivity without compromising the simplicity of the interface. We will also distribute an open implementation of PolemicTweet to facilitate its wider deployment, adoption, and foster improvements.

REFERENCES

1. Barbosa, L., and Feng, J. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, Association for Computational Linguistics (2010), 36–44.
2. Bernstein, M., Brandt, J., Miller, R., and Karger, D. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, ACM (2011), 33–42.
3. Bigham, J. P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R. C., Miller, R., Tatarowicz, A., White, B., White, S., and Yeh, T. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, UIST '10, ACM (New York, NY, USA, 2010), 333–342.
4. Cameron, M. A., Power, R., Robinson, B., and Yin, J. Emergency situation awareness from twitter for crisis management. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, ACM (New York, NY, USA, 2012), 695–698.
5. Senti crowdflower. accessed 6 Sep. 2012, `https://senti.crowdflower.com/senti`.
6. Diakopoulos, N., Naaman, M., and Kivran-Swaine, F. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on* (oct. 2010), 115 –122.
7. Diakopoulos, N. A., and Shamma, D. A. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, ACM (New York, NY, USA, 2010), 1195–1198.
8. Hardesty, L. Making crowdsourcing easier. accessed on 7. Sep. 2012, `http://web.mit.edu/newsoffice/2012/making-crowdsourcing-easier.html`.
9. Huron, S., Haussonne, Y.-M., Monnin, A., and L'hour, Y.-M. Demo: Polemical video annotation by twitter. In *Making Sense of Microposts (#MSM2011)* (2011), 19–21.
10. Huron S., Vuillemot R., J.-D. F. Towards visual sedimentation. *Proc. IEEE Symposium on Visual Analytics Science and Technology (Poster Paper, accepted for publication)* (2012). to appear.
11. Jiang, L., Yu, M., Zhou, M., Liu, X., and Zhao, T. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1 (2011), 151–160.
12. O'Connor, B., Balasubramanyan, R., Routledge, B., and Smith, N. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media* (2010), 122–129.
13. Pang, B., and Lee, L. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval 2*, 1-2 (2008), 1–135.
14. Precision and recall. accessed 6 Sep. 2012, `<http://en.wikipedia.org/wiki/Precision_and_recall>`.
15. Quinn, A. J., and Bederson, B. B. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, ACM (New York, NY, USA, 2011), 1403–1412.
16. Samuel Huron, Raphael Velt, Y.-M. H. Polemic tweet. accessed on 6. Sep. 2012, `http://www.polemictweet.com`.
17. Shipman, F., and Marshall, C. Formality considered harmful: Experiences, emerging themes, and directions on the use of formal representations in interactive systems. *Computer Supported Cooperative Work (CSCW) 8*, 4 (1999), 333–352.
18. Speriosu, M., Sudan, N., Upadhyay, S., and Baldridge, J. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*, Association for Computational Linguistics (2011), 53–63.
19. Tang, A., and Boring, S. #epicplay: crowd-sourcing sports video highlights. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, CHI '12, ACM (New York, NY, USA, 2012), 1569–1572.
20. Willett, W., Heer, J., and Agrawala, M. Strategies for crowdsourcing social data analysis. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, CHI '12, ACM (New York, NY, USA, 2012), 227–236.