Automatic Identification of Personal Life Events in Twitter

Thomas Dickinson
Knowledge Media Institute
Open University, UK

Miriam Fernandez
Knowledge Media Institute
Open University, UK

Lisa A Thomas
Northumbria University
Newcastle upon Tyne, UK

Paul Mulholland
Knowledge Media Institute
Open University, UK

Pam Briggs
Northumbria University
Newcastle upon Tyne, UK

Harith Alani
Knowledge Media Institute
Open University, UK

ABSTRACT

New social media has led to an explosion in personal digital data that encompasses both those expressions of self chosen by the individual as well as reflections of self provided by other, third parties. The resulting Digital Personhood (DP) data is complex and for many users it is too easy to become lost in the mire of digital data. This paper studies the automatic detection of personal life events in Twitter. Six relevant life events are considered from psychological research including: beginning school; first full time job; falling in love; marriage; having children and parent's death. We define a variety of features (user, content, semantic and interaction) to capture the characteristics of those life events and present the results of several classification methods to automatically identify these events in Twitter.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.2.8 [**Database Management**]]: Database Applications—*Data Mining*

General Terms

Social Media, Personal Events

1. INTRODUCTION

For a number of years now, social media technologies have provided us with the means to generate online identities. This has led to an explosion in personal digital data encompassing expressions of self chosen by the individual as well as reflections of self provided by third parties. Increasingly, we are being offered the means of capturing and curating these identities but for many users it is too easy to become lost in the mire of digital data. This is unfortunate as these digital representations of self are not only valuable to the individual as a means of self-reflection, but also have great information value as a community, business and social policy resource.

While a wide body of research in social media has focused on event detection around global events, such as news stories [6] and earthquakes [5], few works have focused on the detection of personal life events.

The automatic detection of personal life events in social media is still a relatively new research topic with the main body of work focusing on classifying tweets about one or two different types of personal events [2], [1].

This paper aims to provide a step forward in this direction by studying the automatic identification of personal life events in Twitter. Unlike Li et al[4] we based the selection of events on previous work from psychological literature, where Jansen and Rubin [3] identified a shared set of life events learned from cultural experience. This work highlights six events that were always present in the top seven most mentioned, including: *beginning school; first full time job; falling in love; marriage; having children and parent's death.* Based on this research our work tries to automatically identify these personal life events in Twitter.

2. DATA COLLECTION, ANNOTATIONS, AND FEATURES

2.1 Data Collection

Our target events are six common life events identified in psychology[3]: Getting Married(GM), Having Children(HC), Starting School(SC), First Full Time Job (FTJ), Death of a Parent(DoP), and Falling in Love(FiL).

In order to seed our initial dataset, we constructed several queries to search Twitter. We used a combination of WordNet, slang, and tense dictionaries to generate them. We also suffixed with "lang:en" to help select only English written tweets.

We set an extraction limit of 1 million tweets per life event, splitting this limit evenly amongst the total queries available per event. After extraction, we ended up discarding the life event "first job" due to insufficient tweets. This is most likely due to the limited number of related terms chosen for that topic.

2.2 Annotations

To annotate our final dataset, we decided to use Crowdflower as our annotation tool. Our questions were:

Table 1: Classifier Results

		J48			NB		
Dataset	Best Features	P	r	F1	P	R	F1
Binary All	sem + ng	0.754	0.753	0.753			
Death of a Parent	int + ng	0.921	0.920	0.920			
Having Children	ng + sem				0.919	0.915	0.915
Getting Married	ng				0.914	0.914	0.914
Starting School	ng				0.934	0.929	0.928
Falling in Love	us+sem+ng				0.853	0.842	0.841

Q1 - Is this tweet related to a particular topic theme?
Q2 - Is this tweet about an important life event?

In the case of Q2, we provided a list of example events taken from Jansen and Rubin's work [3]. Each tweet was annotated by at least three workers. Confident scores are automatically computed by Crowdflower, and return an aggregated result for the annotation based on the responses with the greatest confidence.

From our annotated dataset of 14k tweets, 23% were about events, while 38% were related to the given event theme. This gave us a total of 2241 tweets where we found an intersection between those that were about an event and their target theme. Most event categories have the same amount of tweets, although Falling in Love does have far fewer. This might have been caused by the breadth of our initial root concept, as "love" can cover a wide variety of different topics.

2.3 Features

We split our feature set into four separate feature areas as outlined below:

- *User features*: user features describe the author of the post as well as her standing and participation on the social media platform.
- *Content features*: content features define the vocabulary of the post that its being shared (i.e., the words that compose it) as well as quality measures of the posted text.
- *Semantic features*: semantic features represent the entities and concepts (*Persons*, *Organisations*, *Locations*, etc.) appearing within the post.
- *Interaction features:* interaction features are a novel set of features that look at the network of users who interact with a particular tweet.

3. EXPERIMENT AND RESULTS

3.1 Experimental SetUp

Our training dataset was constructed from our Crowdflower annotations where constructed a 50/50 split of positive vs negative events based on random sampling. After balancing the datasets, we then constructed each post's instance features using the features described in Section 2.3. This resulted in a vector representation of each post with more than 15,000 elements, most of them content and semantic features. For each post we also map its created instance to its class label extracted from the CrowdFlower annotation process (Section 2.3), with 0 denoting the negative class (non event in the case of the event vs. non event classifier and non event of a particular type in the case of the event type classifiers) and 1 denoting the positive class.

For our experiment, we used two different classifiers to compare which worked the best with our datasets: J48, Naive Bayes (NB). We trained each classifier using all permutations of feature sets (e.g., only content features, content features + semantic features, content features + semantic features + user features, etc.). We use 10-fold cross validation to evaluate each of the created machine learning classifiers. We use standard classification performance measures of precision, recall, and F1 measure to assess the performance.

3.2 Results

Table 1 shows the top performing combination of features and classifier for each dataset. As can be seen from the table n-grams are the dominant feature in all cases. Compared with the other results we obtained, we found that few other feature sets made as big of an impact as n-grams. In most cases where n-grams performed alongside other feature sets such as semantics, interaction, and user features, we found most of the gains in performance were minimal. Only Falling in Love seemed to provide a decent gain in performance.

4. REFERENCES

[1] S. Choudhury and H. Alani. Personal life event detection from social media. 2014.

[2] B. D. Eugenio, N. Green, and R. Subba. Detecting Life Events in Feeds from Twitter. *2013 IEEE Seventh International Conference on Semantic Computing*, pages 274–277, Sept. 2013.

[3] S. M. Janssen and D. C. Rubin. Age effects in cultural life scripts. *Applied Cognitive Psychology*, 25(2):291–298, 2011.

[4] J. L. Li, A. Ritter, C. Cardie, and E. H. Hovy. Major life event extraction from twitter based on congratulations/condolences speech acts. In *EMNLP*, 2014.

[5] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. *... of the 19th international conference on ...*, 2010.

[6] C. L. Wayne. Topic detection and tracking in english and chinese. In *Proceedings of the Fifth International Workshop on on Information Retrieval with Asian Languages*, IRAL '00, pages 165–172. ACM, 2000.