Identifying Prominent Life Events on Twitter

Conference or Workshop Item

Identifying Prominent Life Events on 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

Social media is a common place for people to post and share digital reflections of their life events, including major events such as getting married, having children, graduating, etc. Although the creation of such posts is straightforward, the identification of events on online media remains a challenge. Much research in recent years focused on extracting major events from Twitter, such as earthquakes, storms, and floods. This paper however, targets the automatic detection of personal life events, focusing on five events that psychologists found to be the most prominent in people lives. 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. Our proposed classification methods obtain results between 0.84 and 0.92 F1-measure for the different types of life events. A novel contribution of this work also lies in a new corpus of tweets, which has been annotated by using crowdsourcing and that constitutes, to the best of our knowledge, the first publicly available dataset for the automatic identification of personal life events from Twitter.

Categories and Subject Descriptors

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

General Terms

Social Media, Life Events

1. INTRODUCTION

Billions of social media users nowadays post about their daily lives to friends and followers. While a wide body of research in social media has focused on event detection around global events, such as news stories [13], earthquakes [11], or music gigs [8], few works have focused on the detection of common prominent life events,

The automatic detection of prominent 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 life events [4], [3]. However, the benefits for this line of research are numerous. Besides the more obvious areas such as profiling, marketing, and product recommendations, more novel use cases have started to emerge in recent years with the introduction of brief automated biographies (BABs), such as Museum of Me (MoM),[1] and Facebook Lookback (FL).[2] While still in their infancy with regards to how they select content, these tools are examples of ways users can recollect their digital life-log data.

This paper aims to provide a step forward in this direction by studying the automatic identification of prominent life events in Twitter. While other work has looked at identifying a taxonomy of life events automatically[7] we base our research on previous work from psychology literature, where Jansen and Rubin [6] identified a set of 48 common life events. Out of these events, we select the top five, which are *having children*, *beginning school*, *marriage*, *parent's death*, and *falling in love*. Our research here is aimed at automatically identifying these top life events in Twitter. In performing this study we make the following contributions:

- We demonstrate how a collection of user, content and semantic features, that are commonly used in social media analysis, can be applied in the identification of prominent life events
- We study a novel set of interaction features. These features consider the historical interactions among users in order to determine whether unusual patterns of contributions towards a post may indicate that the post refers to a specific life event
- We test a machine-learning approach to automatically identify five prominent life events from psychology, and we evaluate the role of different subset of features on characterising these events. We contrast our findings with existing studies on life events in social media
- We generate and make publicly available a new corpus of 2,241 tweets, manually annotated through Crowd-

KCAP 2015, October 07-10, 2015, Palisades, NY, USA
 ISBN 978-1-4503-3849-3/15/10 ...$15.00
DOI: http://dx.doi.org/10.1145/2815833.2815845.

[1] `http://www.intel.com/museumofme/r/index.htm`
[2] `https://www.facebook.com/lookback`

Flower.[3] To the best of our knowledge, this is the first publicly available dataset for life event identification in Twitter

2. STATE OF THE ART

Event detection using social media is not a new research area. Plenty of work has gone into extracting events from social streams using various methodologies. Sakaki et al [11] use Twitter to identify earthquake by treating tweeters as sensors, and tweets as readings. Jackoway et al[5] identify live news events using Twitter by identifying reliable Twitter users who tweet about such topics. Weng and Lee[14] develop a system called EDCoW that builds signals for individual words, then using wavelet analysis and modularity-based graph partitioning to clustering signals together. Liu et al [8] look at identifying gigs by considering the frequency of images posted to Flickr, within a bounding box around known gig venues.

In closer relation to our work at extracting life events, less research has been done. Eugenio et al [4] looked at identifying two types of life events: marriage, and employment. Their methodology focused on utilising only unigrams, tested over a variety of different classification algorithms. They also considered features like retweets, but found little improvement over unigrams. More recently, Li et al tackled three major challenges with life event detection in twitter: taxonomy of life events, noiseness of data, and lack of training data. They approached this by seeding an initial dataset with tweets that have common key replies like "congratulations", then used an LDA-Clustering & human identification method to construct their training set. They then looked at using an SVM classifier to compare a bag of words, NER, and POS features. Our paper extends the state of the art by investigating six life events that psychologists identified as the most common amongst people.

3. DATA COLLECTION AND ANNOTATION

3.1 Selecting Life Events

As mentioned earlier, our selected life events are those common five identified by Jansen and Rubin [6]: Getting Married(GM), Having Children(HC), Starting School(SC), Death of a Parent(DoP), and Falling in Love(FiL).

3.2 Twitter data collection

In order to seed our initial dataset, we decided to construct several queries for each of our life events. One particular challenge was to reduce bias with the selection of query words. For example, for getting married, using only query words "wedding", "marriage",and "church" might omit a large number of tweets that use different tenses or synonyms. Our approach looked at splitting each event into a minimum number of concepts, and then use WordNet to find related terms. In addition to this, we also extracted slang phrases using OnlineSlangDictionary[4] and the different tenses for each verb with Verbix[5]. This then created a set of words per root concept that we could permutate with each other to generate our final queries. For each query generated we also suffixed with "lang:en" to help select only English written tweets.

Table 1: Life Event root concepts

Life Event	Concepts
Having Children	child, birth
Getting Married	marriage
Death of a Parent	death, parent
Starting School	start, school
Falling in Love	love

To collect data for our study, we took advantage of Twitter's recent indexing infrastructure upgrade. By writing a scraper capable of parsing Twitter's front-end web search, this allowed us to search for tweets with no limit on date. We then used Twitters lookup API to extract all standard meta data associated with each tweet.

We set an extraction limit of 1 million tweets per life event, splitting this limit evenly amongst the total queries available per event. MongoDB was used to store the initial datasets.

3.3 Data Annotation

To annotate our final dataset, we used Crowdflower[6] as our annotation tool. Crowdflower is an online crowd sourcing annotating platform, where uploaded datasets are accompanied by questions for the crowd to make judgements on.

Our goal was to label each tweet with two pieces of information: Is this tweet about an event, and is this tweet about our proposed theme category (e.g., Getting Married)? By using this divided approach, our dataset could then have dual purpose as both a general event classifier, and a topic classifier for future work. Intersecting the two sets of answers where both are "yes", we can consider this list to be tweets about our selected life events.

To help obtain good results, we ran several small trials of annotating around 50 tweets at a time to fine-tune our questions. The main issue we found for both questions was subjectivity. In the case of classifying tweets about events, one annotator might consider that going to the shops is an event, while others were confused that this was just a mundane action. Given our very general definition that an event is something happening to someone, at some time, the first user is correct, however it is easy to see why many users would argue this is not the case. For answering if the tweet is about a particular topic theme, we found that many users would only answer yes if it was only explicit. For example, tweets that were about getting married from a wedding photographers perspective were incorrectly being classified as not being about "getting married". In some cases, people were also getting confused about whether or not the tweet needed to be an event as well as related to the theme. To avoid these problems which appeared during annotation trials, we refined our final CrowdFlower questions as:

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 [6]. By giving this extra guidance to the user, we found it eliminated a lot of the subjectivity we had previously experienced. Examples

[3] `http://www.crowdflower.com/`
[4] http://onlineslangdictionary.com/thesaurus/
[5] http://www.verbix.com/
[6] http://www.crowdflower.com/

of the annotations can be seen in Table 2. Each tweet was annotated by at least three workers. Confident scores are automatically computed by Crowdflower, which returns an aggregated result for the annotation based on the responses with the greatest confidence.[7]

3.4 Generated Dataset

From our annotated dataset of 14k tweets, 23% where 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. Table 4 shows the number of tweets annotated as life events with their respective answers to Q1. 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. This dataset has been made publicly available under `http://reellives.net/rl-data/uploads/2015/06/a692044.csv`

Table 3: Number of tweets annotated as life events in CrowdFlower

	Is This Related To Theme		
Event Type	No	Yes	Total
Death of a Parent	116	509	625
Falling in Love	64	114	178
Getting Married	22	709	731
Having Children	43	489	532
Starting School	51	420	471
Total	296	2241	2537

4. FEATURE ENGINEERING

Understanding how life events are described on the Social Web requires an exploration of the various factors that could characterise these events before identifying which factors are actually important. To inform our understanding of how life events are generated we explore the affects of various features. There are a variety of works that have examined different features in social media to characterise social and content dynamics [9], [2] [12] . However, not all these features may help to represent life events. In this section we list the three different categories of features, and the individual features for each category, that we have selected, including: *user features*, *content features*, and *semantic features*. In addition, in this work we propose a novel set of features; *interaction features*, based on the historical interactions of users within the platform.

- *User features*: user features describe the author of the post as well as her standing and participation on the social media platform, for instance by measuring the user's social connectivity or the frequency of her participation. The hypothesis when using this set of features is that certain types of users (e.g., those that are more popular, or followed by more users within the network) may be more prone to share life events in Twitter.
- *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, such as metrics to calculate the readability of the post or the post's sentiment. The hypothesis behind using these features is that posts related to life events may be written in a different way than more general posts (e.g., express stronger sentiment, positive or negative, be written in a more formal/colloquial way, etc.)
- *Semantic features*: semantic features represent the entities and concepts (*Persons*, *Organisations*, *Locations*, etc.) appearing within the post. The hypothesis for using these features is that prominent life events may be semantically associated with certain entities or concept types.
- *Interaction features:* interaction features are a novel set of features that look at the network of users who interact with a particular tweet. Rather than just consider number of retweets, favourites or replies, we consider who are the users performing these actions and their interaction patterns towards the author of the post. The hypothesis for using these features is that, if users that do not generally interact with the author of the post they suddenly show interest, the post may contain information of special interest.

The list of individual features considered as part of every feature set are specified in Tables 4, 5, 6 and 7.

5. LIFE EVENT CLASSIFICATION

Once the set of features have been selected our goal is to assess how posts about life events differ in terms of the selected factors. To do so, we utilise a two-stage approach that functions as follows:

1. *Identify life events:* we first detect those posts that talk about a life event vs. those posts that don't.
2. *Identify particular types of life events:* in a second step we detect which posts talk about a particular event type.

Details of the experimental setup are explained in the following section.

5.1 Experimental SetUp

To uncover the factors that characterise life events, as well as particular types of life events, we first derive the set of post that will constitute the instances to train a machine learning classifier. As described in Section 3.4, the CrowdFlower annotation resulted in large class imbalances between posts that refer to a particular type of event vs. posts that don't. In all occasions there are more posts that refer to events of the specific type (i.e., instances of the positive class) than to instances of the negative class. In order to ensure that we have a balanced class distribution in each dataset, we randomly sample tweets categorised as non-events to even the dominant class (in our case the positive class) from each respective dataset. This resulted in a 50:50 split between posts referring to a life event type vs. posts that don't. The resultant number of instances in each dataset is shown in Table 8.

After balancing the datasets, positive vs. negative class distribution, we then constructed each post's instance features using the features described in Section 4. This resulted in a vector representation of each post with more than 15,000

[7] `https://success.crowdflower.com/hc/en-us/articles/202703305-Getting-Started-Glossary-of-Terms`

Table 2: Example answers for the event category Getting Married

Tweet	*Q1*	*Q2*
The Patriots obvi won because Im getting married tomorrow and they wanted me to be in a good mood.	Yes	Yes
not my idea of marriage	Yes	No
This superstar fianc sang the flood to me while I was giving birth!	No	Yes
I'm sorry for the rant, but I feel strongly about this. & I wish a celebrity like you could stand up and espouse this cause	No	No

Table 4: User Features

In-degree	For the author of each post, this feature measures the number of incoming connections to the user. Very popular users may be more keen on sharing their prominent life events than users with smaller networks
Out-degree	This feature measures the number of outgoing connections from the user. Users who follow a big network tend to listen more than speaking (i.e., posting), and therefore may be less keen on sharing their own life events
Post Count	Measures the number of posts that the user has made during her life in the social networking platform. Users that post, i.e., share more, may be more keen on sharing prominent life events.
User Age	Measures the length of time that the user has been a member of the community. The longer a user has been participating in the platform, the higher the probability that he/she has experienced an important life event during this period. Therefore, users that have been part of the community for longer time may have shared more life events
Post Rate	Measures the number of posts made by the user per day. The hypothesis is that users who share more frequently, they may also feel more comfortable with sharing prominent life events than those that posts with less frequency
Country	This feature describes the user's country as stated by him in his social media profile. Note that, for the purpose of this investigation, seven english-speaking countries have been considered (Canada, Australia, Singapore, United States, Indonesia, Ireland and Great Britain). The hypothesis behind considering this feature is that culture may play a role on how keen the users are on sharing life events

Table 8: Instances per Class

Instance Count	Event (Type)	Non Event (Type)
Event	2537	2537
Death Parent	509	509
Having Children	489	489
Getting Married	709	709
Starting School	420	420
Falling in Love	114	114

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 3.4), 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.

Each dataset is used as input of three different classifiers J48, Naive Bayes (NB) and Support Vector Machines (SVM). 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. We report the results obtained for each dataset for the J48 and Naive Bayes classifiers in Section 5.2. We left out the results obtained with SVM for clarity, since it is the classifier that obtained worst performance across datasets. Results are reported for each feature set individually as well as for the optimal combination of features (last row of each table).

5.2 Results

We begin by examining the performance of different feature sets on identifying posts about life events. Table 9 presents the performance that the J48 and Naive Bayes classifiers achieve when trained on isolated feature sets as well as the best performance combination of features. We note that, for the isolated feature sets content features are the best performing features obtaining a 0.738 F1-measure. Among content features, the most discriminative ones are n-grams. The top classifier is the J48 using semantic + n-gram features, obtaining 0.753 F1-measure. By performing attribute selection over content features using information gain we identify some of the most discriminative n-grams including: *knot* and *tied*, which refer to the metaphorical expression *tie the knot*, *passed* and *birth*, which refer to the start and end of life and terms such as *love*, *loving* and *loved*, which indicate affection. Among the most discriminative semantic features we find `http://en.wikipedia.org/wiki/Knot`, `http://en.wikipedia.org/wiki/Wedding`, `http://en.wikipedia.org/wiki/Pillow` and `http://en.wikipedia.org/wiki/Hessian_(cloth)`, which refer to the dressing code and decorations usually used in weddings and, `http://en.wikipedia.org/wiki/Funeral` and `http://en.wikipedia.org/wiki/Cancer`, which were associated with the "death of a parent" events.

The rest of the tables show the results for the identification of particular event types. As we can see in these tables content features are the best performing individual features in all cases, sometimes using the J48 and sometimes using the Naive Bayes classifier. Among content features the most discriminative ones are n-grams followed by polarity. Interaction features, as opposed to our initial hypothesis, are not very useful discriminating life events or event types. This may be due to the fact that Twitter may not be the platform

Table 5: Content Features

nGrams	nGrams represent the vocabulary used in the tweets, which may help to discriminate different types of prominent life events. In the case of this work we have considered unigrams or unique terms. A total of 29334 unigrams are identified in the collected corpus. To reduce the sparsity of the vocabulary, following the approach of Saif [10] we have removed all those infrequent terms appearing only once, remaining with 7948 unigrams.
Post Length	Number of words in the post. Although tweets allow a maximum of 140 characters, longer/shorter messages may be associated to different types o life events
Complexity	Measures the cumulative entropy of terms within the post to gauge the concentration of language and its dispersion across different terms. Let n be the number of unique terms within the post p and f_i is the frequency of term t within p, therefore complexity is given by: $$\frac{1}{n}\sum_{i=1}^{n} f_i(\log n - \log f_i) \quad (1)$$ This feature aims to study whether posts describing life events contain many terms which are not repeated often or rather repeat terms from a limited vocabulary
Readability	Gunning fog index using average sentence length (ASL) and the percentage of complex words (PCW), i.,e., those with 3 or more syllables [8]: $0.4 * (ASL + PCW)$ This feature gauges how hard the post is to parse by humans.
Referral Count	Count of the number of hyperlinks within the post (i.e., links to additional information). Users may provide external references when it comes to describe their life events
Time in Day	The number of minutes through the day that the post was made. This feature is used to identify possible key times within the day associated with life events.
Informativeness	The novelty of the post's terms with respect to other posts. We derive this measure using the Term Frequency-Inverse Document Frequency (TF-IDF) measure. Posts about prominent life events may contain unique terms with respect to the platform's vocabulary or rather terms that are familiar to the platform's users. $$\sum_{t \in p} tf_{t,p} \times idf_t \quad (2)$$
Polarity	Assesses the average polarity of the post (positive, negative, neutral) using SentiStrength.[9]. Posts about important life events may be associated with stronger (positive or negative) sentiment.
Mentions	Count of the number of mentions to other users within the post. When announcing particular life events the author of the post may mention the relevant people involved.
Num Retweets	Count of the number of times the post has been shared (re-posted) by other users. Posts about prominent life events may be less retweeted, since other users/friends may feel cautious about sharing the author's personal information

Table 9: Binary All

	J48			NB		
Features	P	R	F1	P	R	F1
interaction	0.567	0.567	0.566	0.519	0.505	0.393
user	0.574	0.572	0.569	0.525	0.508	0.406
content	0.738	0.738	0.738	0.726	0.721	0.712
semantic	0.593	0.593	0.592	0.592	0.588	0.583
sem + ng	0.754	0.753	0.753			

of choice for people to post about personal life events. Twitter allows users to post 140 character messages (tweets) and follow the messages of other users on their Twitter feed. It is mainly used to communicate with other individuals with similar interests, regardless of whether users know one another, and to follow updates from celebrities, companies, etc. On the other hand, platforms like Facebook are mainly used by individuals who wish to stay connected with, or reconnect with, people that they know offline. In this sense, it is more likely that users share their life events on platforms like Facebook.

However, while Twitter information is mostly public, obtaining information from Facebook requires the explicit consent of every single user from whom information is collected. This constitutes a strong barrier to collect data from this platform. We are currently working on developing an application that will allow us to explore this platform's information as part of our future line of work.

Table 10: Death of a Parent

	J48			NB		
Features	P	R	F1	P	R	F1
interaction	0.59	0.589	0.588	0.57	0.548	0.509
user	0.638	0.632	0.628	0.557	0.517	0.413
content	0.913	0.913	0.913	0.919	0.916	0.915
semantic	0.62	0.62	0.619	0.633	0.633	0.632
int + ng	0.921	0.920	0.920			

Table 10 shows the results obtained when trying to automatically identify posts about the death of a parent. As we can see by the results, the most powerful individual features

Table 6: Semantic Features

Entities	This feature set represents the entities that appear in the collected Twitter corpus. A total of 7373 unique entities have been identified by using the entity extractor API TextRazor[10]. Examples include: `http://en.wikipedia.org/wiki/MasterCard`, `http://en.wikipedia.org/wiki/Breastfeeding` , `http://en.wikipedia.org/wiki/High_school`
Concepts	This feature set represents the concepts (or entity types) extracted from the collected Twitter corpus. 163 unique concepts have been identified by looking up each entity's `rdf:type` in the DBPedia ontology and recording these concepts in a list for each post. Examples include: `http://dbpedia.org/ontology/Work`, `http://dbpedia.org/ontology/Food`, `http://dbpedia.org/ontology/Hospital`

Table 7: Interaction Features

Naive Interaction	Total number of users $\|U\|$ who interact (retweeted or reply) with the tweet.
Interaction Ratio	Lets be U be the set of users who have interacted (retweeted, contributed to conversation) with the tweet, and a the author of the tweet. The interaction ratio considers the number of times that the user $u_i \in U$ mentions the author of the post a in his previous conversations $m(u_i, a)$ with respect to all his mentions to other users $m(u_i)$. $$\sum_{i=1}^{n} \frac{m(u_i, a)}{m(u_i)} \quad (3)$$
Length Conversation	This feature counts the number of days a Twitter conversation lasts. Those conversations that last several days might be an indicator of an event, for example, a number of people saying congratulations to someone's post about getting married

are content features, and within them, n-grams, obtaining up to 0.915 F1-measure with the NB classifier. By doing attribute selection using information gain, the top 10, most discriminative terms about this event are: *passed*, *pass* and *expired*, which refer to the act of death, *father*, *mother* and *dad*, which identify the person, *loving* and *sad*, which refer to the feelings involved in this type of life event. Negative sentiment associated to the post is also a discriminative feature of this type of event. For this particular event type, the combination of n-grams and interaction features obtains the best results (0.92 F1- measure), providing a slightly better performance that content features alone. Our hypothesis is that, in the context of death, users who do not tend to interact are more sympathetic and explicitly express their condolences.

Table 11: Having Children

	J48			NB		
Features	P	R	F1	P	R	F1
interaction	0.592	0.591	0.589	0.57	0.57	0.57
user	0.608	0.606	0.605	0.593	0.547	0.484
content	0.897	0.897	0.897	0.918	0.914	0.914
semantic	0.586	0.582	0.577	0.617	0.614	0.613
co + sem				0.919	0.915	0.915

Table 11 shows the results obtained for recognising the life event of having children. Again in this case, content features are the best performing individual set of features, obtaining up to 0.914 F1 measure with the NB classifier. The most discriminative features are n-grams and among them: *birth*, *gave*, *born* and *delivered*, which refer to the act of having a child, *baby*, *child* and *school*, which refer to the child growing up, and *loves* and *love*, which express the feelings experienced with this type of life event. Other discriminative features include positive polarity and a higher number of referral counts or references to other users (the user who generates the post tends to mention his/her partner in the post). The top combination of features are content and semantic features, but the improvement is only marginal. Among the most discriminative semantic features we can highlight: `http://en.wikipedia.org/wiki/Infant`, `http://en.wikipedia.org/wiki/Childbirth` and `http://en.wikipedia.org/wiki/Wife`

Table 12: Getting Married

	J48			NB		
Features	P	R	F1	P	R	F1
interaction	0.575	0.57	0.563	0.555	0.518	0.419
user	0.577	0.572	0.565	0.613	0.524	0.407
content	0.907	0.907	0.907	0.898	0.898	0.898
semantic	0.698	0.686	0.681	0.784	0.688	0.659
ng				0.914	0.914	0.914

Table 12 shows the results for the identification of the event getting married. For this event, content features are again the top individual set of features. The top 10 identified n-grams that characterise this life event are: *knot*, *tied*, *tying* and *tie*, which refer to the expression *tie the knot*, *wedding*, *ceremony*, *wedded*, *marriage* and *married*, which refer to the act of getting married. N-grams alone obtain an F1 measure of 0.914, and adding other subsets of features does not boost this performance.

Table 13 shows the results for the identification of the starting school event. Content features are the top individual features, obtaining an F1 score of 0.925. Among the most discriminative n-grams for this life event are: *school*, *staring*, *start*, *started*, *tomorrow*, *beginning*, *begin* and *year*, which refer to the event of starting school and *loved*, *loves*

Table 13: Starting School

	J48			NB		
Features	P	R	F1	P	R	F1
interaction	0.537	0.537	0.537	0.565	0.551	0.527
user	0.581	0.58	0.578	0.564	0.564	0.564
content	0.924	0.923	0.922	0.931	0.925	0.925
semantic	0.599	0.593	0.587	0.609	0.605	0.601
ng				0.934	0.929	0.928

which refers to loving the first day at school. Informativeness and complexity are also relevant content features, indicating that posts about starting school tend to always use the same vocabulary, which are well known terms to the platform's users.

Table 14: Falling in Love

	J48			NB		
Features	P	R	F1	P	R	F1
interaction	0.473	0.474	0.469	0.53	0.518	0.461
user	0.491	0.491	0.489	0.542	0.522	0.457
content	0.769	0.767	0.767	0.752	0.737	0.733
semantic	0.469	0.469	0.469	0.564	0.561	0.556
us+sem+ng				0.853	0.842	0.841

Table 14 shows the results for the identification of the starting school event. Content features are the top individual performing feature set, obtaining an F1 measure of 0.767. This result is improved up to F1 measure 0.841 by including semantic and user features. The key n-grams identified for this event are *love*, *happy*, *loved*, *amazing*, *friend*, *starting*, and *girlfriend*. Other relevant content features include polarity, which tends to be positive. Top semantic features include `http://en.wikipedia.org/wiki/Girlfriend`, and `http://en.wikipedia.org/wiki/Music_download`.

6. DISCUSSION

One of the most fascinating dimensions of social media is the way in which it allows us to create our digital lifelogs and help us to maintain an autobiographical memory. Technology makes it possible to identify, retrieve, or relive parts of those important life events, and many social media platforms are developing tools to this effect. In this paper we focused on the automatic identification of prominent life events from Twitter data, considering those events from psychology research that were identified as the most important events that people may experience is their lifetime.

Our study focused on how users share life events on Twitter. However, as one would expect, the obtained results may vary across different social platforms. We indeed observed from our results that interaction features were not particularly discriminative within the Twitter platform. This is most likely due to the fact that ego-networks on Twitter tend to expand well beyond close friends and family, and thus social interaction patterns do not appear to change much around event-specific posts. On the other hand, recent studies suggested that Facebook is generally used to disclose important life events. We therefore aim to explore Facebook as part of our feature work [1], where we expect to see a greater impact of interaction features on life event detection.

Our selection of content, semantic and user features was inspired by the literature [9], [2] and [12], and consisted of features that we found relevant, in one aspect or another, to identify life events. However, we acknowledge the fact that there could be other features that might better characterise our selected five events for individual platforms, and across platforms.

Regarding our proposed set of interaction features, note that at the moment Twitter does not allow the collection of the historical chain of favourites, replies and retweets towards a user. We therefore generated our interaction features based on mentions. However, it will be desirable to test the proposed interaction features with all the above interaction types.

Regarding the generated dataset, one could also argue that, despite the use of WordNet to expand the list of life event root concepts, tweets were selected based on keywords rather than randomly, and therefore n-grams may have an advantage as individual features to identify these events. A random collection of tweets would however not ensure gathering tweets about the life events we aimed to study, and would have required far more data annotations in order to obtain a similar sample, thus significantly increasing the annotation costs.

As mentioned in the paper, we generated a gold standard dataset of life event annotations using CrowdFlower. We now believe that the annotations quality could be improved by providing additional context to the annotators instead of only showing them the individual post. It is a fairly difficult task, even for a human annotator, to decide whether a tweet is about an event or not, without any context. For example, the tweet *MadJacks Forever Memories* conveys little information to deduce that it talks about getting married at a casino called MadJacks in Vegas. In addition to this, our selection method extracts any tweet that has been found to include our original query words. This may possibly cause a problem where a tweet is taken out of context from the original.

We noted in the paper that we experimented with various questions to provide to the CrowdFlower annotator. Next we plan to manually evaluate the CrowdFlower annotations, to reassess their quality and to identify where difficulties and confusion might have emerged. This will enable us to further improve these manual annotations to ensure a higher quality gold standard and results.

With regards to the life events we studied in this paper, although we based our selection of events on psychology research, it can be argued that some of these events might be difficult to discover in social media. For example, with the event "falling in love" we found many tweets in our seed dataset. However, when annotated, very few of these were actually about falling in love. This might be because our initial query words were inadequate, but it could also be due to lack of sufficient information in single tweets to determine whether someone has fell in love with another person.

This paper only studies the top five events that were deemed to be most popular by Jansen and Rubin [6]. It would be interesting to expand our study to more, or all the 48 specific events listed in that research. However, some of those events might be *difficult* to find on social media, such as "Own death", whereas others might only be found

in reminiscing posts, such as the "Own birth" event.

In summary, while there is still extensive room for future work, our experiments and results show that different features can be used to accurately identify personal life events in social media, helping therefore to manage our Digital Personhoods. We hope that the presented study will serve as a basis for future work within the community and enable further research into the automatic identification and management of our digital identity.

7. CONCLUSIONS

Although much research has been done on identifying global events from social media, very little attention has been given to extracting prominent life events.

This paper is one of the first to tackle this issue, by developing methods to automatically identify six types of life events, which have been found to be amongst the most common and highly memorable events by psychologists. These events are *beginning school; falling in love; marriage; having children* and *parent's death*.

We study in this paper how different *user*, *content*, *semantic* and *interaction* features can be used to characterise and identify these personal life events, and applied a machine learning approach for automatic event identification. Classifiers trained with the proposed features obtain results between 0.84 and 0.92 F1-measure.

We contrasted the role of different features and concluded that content features are the most discriminative features to identify these life events. Negative polarity and a higher interaction ratio are characteristics of the *death of a parent* event. Positive sentiment and references to other users are characteristics of *having children* event posts. Semantic concepts around decorative elements are characteristic of *getting married* posts. Common and simple vocabulary are characteristics of *starting school*; and post about *falling in love* tend to show a positive sentiment. Additionally, we identified the set of n-grams that were the most discriminative for each type of event. We also found almost no effect of the interaction features on life event detection when applied to Twitter.

An additional contribution of this work lies in the generation of a dataset for personal event detection. This dataset has been made publicly available to the community.

8. REFERENCES

[1] J. L. Bevan, M. B. Cummings, A. Kubiniec, M. Mogannam, M. Price, and R. Todd. How are important life events disclosed on facebook? relationships with likelihood of sharing and privacy. *Cyberpsychology, Behavior, and Social Networking*, 18(1):8–12, 2015.

[2] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proc. Int. AAAI Conf on Weblogs and Social Media (ICWSM)*, USA, 2010.

[3] S. Choudhury and H. Alani. Personal life event detection from social media. In *Workshop on Social Personalisation, Hypertext*, Santiago, Chile, 2014.

[4] B. D. Eugenio, N. Green, and R. Subba. Detecting Life Events in Feeds from Twitter. In *IEEE Seventh International Conf on Semantic Computing (ICSC)*, CA, USA, Sept. 2013. IEEE.

[5] A. Jackoway, H. Samet, and J. Sankaranarayanan. Identification of live news events using twitter. In *Proc of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, LBSN '11, pages 25–32, New York, NY, USA, 2011. ACM.

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

[7] J. L. Li, A. Ritter, C. Cardie, and E. H. Hovy. Major life event extraction from twitter based on congratulations/condolences speech acts. In *Proc of the 2014 Conf on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014.

[8] X. Liu, R. Troncy, and B. Huet. Using social media to identify events. *3rd Workshop on Social Media (WSM'11), ACM Multimedia Conf*, 2011.

[9] M. Rowe and H. Alani. Mining and comparing engagement dynamics across multiple social media platforms. In *Proc of the 2014 ACM Conf on Web science*, Bloomington, USA, 2014. ACM.

[10] H. Saif, M. Fernandez, Y. He, and H. Alani. On stopwords, filtering and data sparsity for sentiment analysis of twitter. In *Proc of the 9th Int. Conf on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014.

[11] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proc of the 19th Int. Conference on World Wide Web (WWW)*, NY, USA, 2010.

[12] B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *IEEE 2nd Int. Conf on Social Computing (SocialCom)*, MN, USA, 2010. IEEE.

[13] C. L. Wayne. Topic detection and tracking in english and chinese. In *Proc of the 5th Int. Workshop on on Information Retrieval with Asian Languages*, IRAL '00, pages 165–172, Hong Kong, China, 2000. ACM.

[14] J. Weng and B.-S. Lee. Event detection in twitter. In *Proc. Int. AAAI Conf on Weblogs and Social Media (ICWSM)*, Barcelona, Spain, 2011.