Identifying Important Life Events from Twitter Using Semantic and Syntactic Patterns

IDENTIFYING IMPORTANT LIFE EVENTS FROM TWITTER USING SEMANTIC AND SYNTACTIC PATTERNS

Thomas Dickinson[1], Miriam Fernandez[1], Lisa A Thomas[2], Paul Mulholland[1], Pam Briggs[2] and Harith Alani[1]
[1]KMI Open University
[2]Northumbria University

ABSTRACT

Identifying global events from social media has been the focus of much research in recent years. However, the identification of personal life events poses new requirements and challenges that have received relatively little research attention. In this paper we explore a new approach for life event identification, where we expand social media posts into both semantic, and syntactic networks of content. Frequent graph patterns are mined from these networks and used as features to enrich life-event classifiers. Results show that our approach significantly outperforms the best performing baseline in accuracy (by 4.48% points) and F-measure (by 4.54% points) when used to identify five major life events identified from the psychology literature: Getting Married, Having Children, Death of a Parent, Starting School, and Falling in Love. In addition, our results show that, while semantic graphs are effective at discriminating the theme of the post (e.g. the topic of marriage), syntactic graphs help identify whether the post describes a personal event (e.g. someone getting married).

KEYWORDS

Semantic Networks, Event Detection, Frequent Pattern Mining, Classification, Social Media

1. INTRODUCTION

Social media platforms are not only the online spaces where billions of people connect and socialise, they are also where much of their prominent life events are captured, shared, and reflected on. Most of these platforms are exploring methods that make use of this content for the purpose of profiling, marketing, and product recommendation, which constitutes a really valuable resource for governments and organisations. More recently, several initiatives are also focusing on exploring the value that this content has for the individuals as a means of self-reflection. Services such as Museum of Me, Facebook Look Back, and Google Creations aim to generate brief automated biographies for the users based on their generated content. Although these services and technologies are still in their infancy, they raise the need for a more accurate identification of personal life events.

However, as opposed to the identification of large-scale events from social media (e.g., detection of news stories (Wayne, 2000), natural disasters (Sakaki, et al., 2010), music events (Liu, et al., 2011), etc.), which has gained much interest in recent years, the identification of personal life events (e.g. getting married, having a child, starting school) is still largely unexplored. This problem poses new requirements and challenges, since these events do not receive the same amount of coverage as large-scale events (unless they involve celebrities (Choudhury & Alani, 2014)), and do not show similar geographical or temporal distributions. In addition, the accurate identification of life events in social media requires distinguishing between posts that mention a life event (e.g., I'm getting married in the church today) and posts about the same theme that do not mention a life event (e.g., check out my wedding photography business). While previous works have found that uni-grams are one of the best features for thematically classifying posts in terms of whether or not they mention the topic of a life event (Di Eugenio, et al., 2013) (Dickinson, et al., 2015), current methods cannot adequately distinguish posts describing life events from posts of the same theme that do not.

In our work, we focus on identifying events from Twitter data that are regarded as some of the most significant life events in the psychology literature (Janssen & Rubin, 2011): Getting Married, Having Children, Starting School, Death of a Parent, and Falling in Love. To address the above mentioned challenges we propose a novel approach based on the representation of posts as syntactic and semantic graphs. Our hypothesis is that, by considering the conceptualisations that emerge from social media posts (mined in our work by using WordNet and ConceptNet), as well their syntactic structures (extracted by applying dependency parsing), we can obtain relevant insights for a more accurate detection of personal life events. These insights are extracted by performing frequent pattern mining over the generated graphs and used as features to generate life event classifiers. Our results show that our approach significantly outperforms the state of the art baseline in accuracy (by 4.48% points) and F-measure (by 4.54% points). In addition, we can observe from our results that, while semantic graphs are effective at discriminating the theme of the post, syntactic graphs help identifying whether the personal event is present in the post or not. The main contributions of this paper are:

1. Introduce a semantic and syntactic graph-based approach for identifying five personal-life events using ConceptNet, WordNet, and Dependency Parsing
2. Mine frequent graph patterns, and serialise them as features, in a LibLINEAR classification algorithm
3. Compare several feature reduction techniques to elicit the best performance
4. Construct tri-class classifiers to categorise posts into those about an event, those about the theme of an event, and those that are neither
5. Compare results to the state of the art baseline, showing an average increase of 4.48% points in accuracy, 4.54% points in F-measure
6. Discuss the advantages of using syntactic and semantic graphs for personal life event identification

2. STATE OF THE ART

As mentioned earlier, much work has been done on detecting global and/or famous events, such as natural disasters (Sakaki, et al., 2010), music events (Liu, et al., 2011), and prominent news stories (Wayne, 2000). However, detecting life events differs from the detection of such global events. Life events tend to focus on one particular individual, and only very few posts from a user's history are likely to refer to a particular life event. Few works in the literature have attempted to address this problem. Di Eugenio et al (Di Eugenio, et al., 2013) have investigated the identification of life events related to employment, and marriage. Their approach compared several different classifiers trained with different subsets of linguistic features, concluding that uni-grams are the best features to automatically identify these types of event. Li et al (Li, et al., 2014) took a slightly different approach. Rather than trying to identify a set of pre-defined life events, this approach focused on detecting life events mentioned on Twitter by clustering posts based on congratulatory and condolence replies such as "Congratulations", and "Sorry to hear that". Their approach is also based on the use of linguistic features including sequences of words within the tweets, named entities, and topic models for different life event categories. Choudhury and Alani (Choudhury & Alani, 2014) studied the use of interaction features to identify several events: marriage, graduation, new job, new born, and surgery. Their approach aggregates linguistic features with interaction features (i.e, features based on the interactions of users with their social network - comments, replies, etc.).

Our work extends the above approaches by representing social media posts as graphs, to capture the interdependencies of linguistic, semantic and interaction features. Although using graphs for finding similar concepts is not a new idea (e.g., (Resnik, 1999) and (Rada, et al., 1989)), applying it for identifying life events is vastly under researched. Sayyadi et al (Sayyadi, et al., 2009) used graphs for news-event identification, by creating a "key graph" of key words co-occurring within a document, where nodes represent individual keywords and edges represent their co-occurrence. To the best of our knowledge, our work is the first one which uses a combination of syntactic, and semantic graphs for personal life event identification.

3. APPROACH

3.1 Representing Posts as Feature Graphs

Before we can extract frequent patterns from social media, we first need to represent our posts as graph structures. An example of a full feature graph for the tweet "**I had a baby**" is displayed in Figure 1**.** Our graphs can be placed into two specific groups: Semantic, and Syntactic.

We construct semantic graphs by extracting concepts and synsets from the posts, then expanding into two popular semantic networks: ConceptNet, and WordNet. Our hypothesis is that, after expansion, posts may have related nodes amongst one another. For example, the concepts "mother" and "father" are related by the concept "parent" within ConceptNet. By extending these two posts into ConceptNet, /c/en/parent would then become a feature, appearing in two posts. In this work, we only expand 1 hop into both ConceptNet and WordNet, to limit the size of the graphs we create. In the case of WordNet, we extend one hop using both hypernyms, and hyponyms with an "IS_A" relationship, whereas for ConceptNet, we include all relationship types. As concepts in ConceptNet are normalised WordNet URIs, we extract N-grams from each post, and use ConceptNet's URI standardisation API.[1]

Syntactic graphs can be divided into two areas: dependency graphs, and token graphs. Dependency graphs can be constructed using dependency parsing (Covington, 2001), which are the syntactic dependencies between words within a sentence. Dependency parsers represent these dependencies as acyclic directed trees with a word, generally the main verb, representing the main node, root, of the tree, and the rest of the words being dependants of the head or other governor nodes. These trees constitute a graph, where words and their associated part of speech (POS) are the nodes, and the edges are the syntactic dependencies between these nodes. Our hypothesis is that posts about the same life events may share similar syntactic structures. In this work, we use the Stanford Neural Network Dependency parser (Chen & Manning, 2014) to obtain these dependency trees from which syntactic graphs are generated. Use of dependency graphs and n-grams within event detection is not uncommon (Li, et al., 2014). However, we have not seen another approach that mines frequent sub-graphs from syntactic graphs, in order to create feature sets for life event classifiers. With token graphs we represent the different tokens (words) that emerge from social media posts, with their associated POS, as nodes, and the sequence in which they appear within the posts (i.e., one token before/after another token) as the edges. This is very similar to n-grams. Our hypothesis is that posts about a particular life event may share common sequences of either tokens or POS tags. We use TweetNLP (Owoputi, et al., 2013) to extract both tokens and POS tags from the text of each post.

3.2 Mining Frequent Patterns

Once the previously described graphs have been generated from the social media posts, we apply frequent pattern mining to the graphs associated with each type of life event. By frequent, we mean a sub-graph appearing more than n times within a set of graphs. As the size of our training sets are not very large (table 1), we set n to 2. This allows us to identify as many sub-graphs as possible, without returning a posts entire graph. As this will return a very large set of graphs, with many sharing the same information, we only search for closed subgraphs, where a closed subgraph has no parent with the same frequency.

To mine these frequent sub-graphs, we use CloseGraph, which can perform an exhaustive frequent sub-graph search over our graph data sets, returning all frequent closed graphs. We use the Parallel and Sequential Mining Suites (ParSeMiS[2]) implementation of this algorithm. As this type of mining has a high memory cost, we represent our graph as a cyclic structure, collapsing nodes. In addition, we also limit the maximum number of edges allowed in a sub-graph, reducing the search space.

[1] https://github.com/commonsense/conceptnet5/wiki/API#uri-standardization
[2] https://www2.informatik.uni-erlangen.de/EN/research/zold/ParSeMiS/index.html

ISBN: 978-989-8533-57-9 © 2016

4. EXPERIMENT SETUP

4.1 Dataset

Our dataset was collected back in December 2014, by mining Twitter's advanced search features, using query expansion into WordNet with several key concepts for each type of life event. More information can be found in our previous paper (Dickinson, et al., 2015) This dataset is composed of 12,241 tweets manually annotated through CrowdFlower. Five relevant life events, common across age and culture (Janssen & Rubin, 2011), are represented in this dataset**: Starting School; Falling in Love; Getting Married; Having Children; and Death of a Parent**.

For each tweet, annotations contain answers to two questions: (i) Is this tweet about an important life event? and (ii) Is this tweet related to a particular event theme? **-** where event theme refers to one of the previously mentioned types of life events. Additionally, we introduce a random sample of tweets collected (2000) and internally annotated from a Twitter sample endpoint to generate a "not about event or theme" class. This represents a class where neither category of theme or event exists. The CrowdFlower annotations for this dataset are publicly available[3].

4.2 Training Set Selection

We use the posts of the previously described dataset to generate the feature graphs (see Section 3.1). These feature graphs are then mined, and the extracted patterns are used as features to train classifiers for each type of life event (Death of a Parent, Having Children, Falling in Love, Getting Married, and Starting School). For each life event, we construct a tri-class classifier with the following labels: About Event and About Theme (+E+T), Not About Event and About Theme (-E+T), Not About Event and Not About Theme (-E-T).

As named, +E+T contains those posts that are both about a particular event, and its given theme, while those in -E+T contain posts that are thematically similar, but do not indicate an event has happened. Our third class, -E-T, contains tweets extracted from Twitter's random sample endpoint. In each case, we balance the training set against the size of the smallest class, with the following distributions: Getting Married (706 per class), Starting School (419 per class), Having Children (387 per class), Death of a Parent (116 per class), and Falling in Love (114 per class). To consider if our approach is statistically significant, we generate our dataset 10 times for each life event, so that we can perform t-tests against our baseline.

4.3 Classifier Generation and Evaluation

Each training dataset, and combination of features (semantic ConceptNet, semantic WordNet, syntactic), is used as input to a LibLINEAR classifier. LibLINEAR works particularly well with very large feature sets. We also independently tested several other classifiers (J48, Naive Bayes, and SVMS with linear, polynomial, sigmoid, and radial bias kernels), but in all cases, LibLINEAR outperformed each in both F-Measure, time performance, and accuracy. The generated classifier is tested using 10-fold cross validation considering Precision, Recall and F1-measure as evaluation metrics. In section 5, we report some of our best performing feature sets using this classification and evaluation set-up.

4.4 Baseline Selection

For our baseline, we implement the features outlined in Li et al. These included **N-grams**, **topic dictionaries**, and **entities**. Entities are extracted using TextRazor[4], and for the case of topic dictionaries, our topic models are generated slightly differently to Li et al, where we discover 5 topics (representing our five chosen life events) across our total annotated dataset. The top 40 words for each topic are used as our topic dictionaries, as outlined in Li et al. Rather than treat n-grams, topics, and entities as a single baseline, for each experiment, we run our classifiers against every permutation of the baseline features, and select the best one to compare against our graph approach.

---

[3] Link removed for blind submission
[4] https://www.textrazor.com/

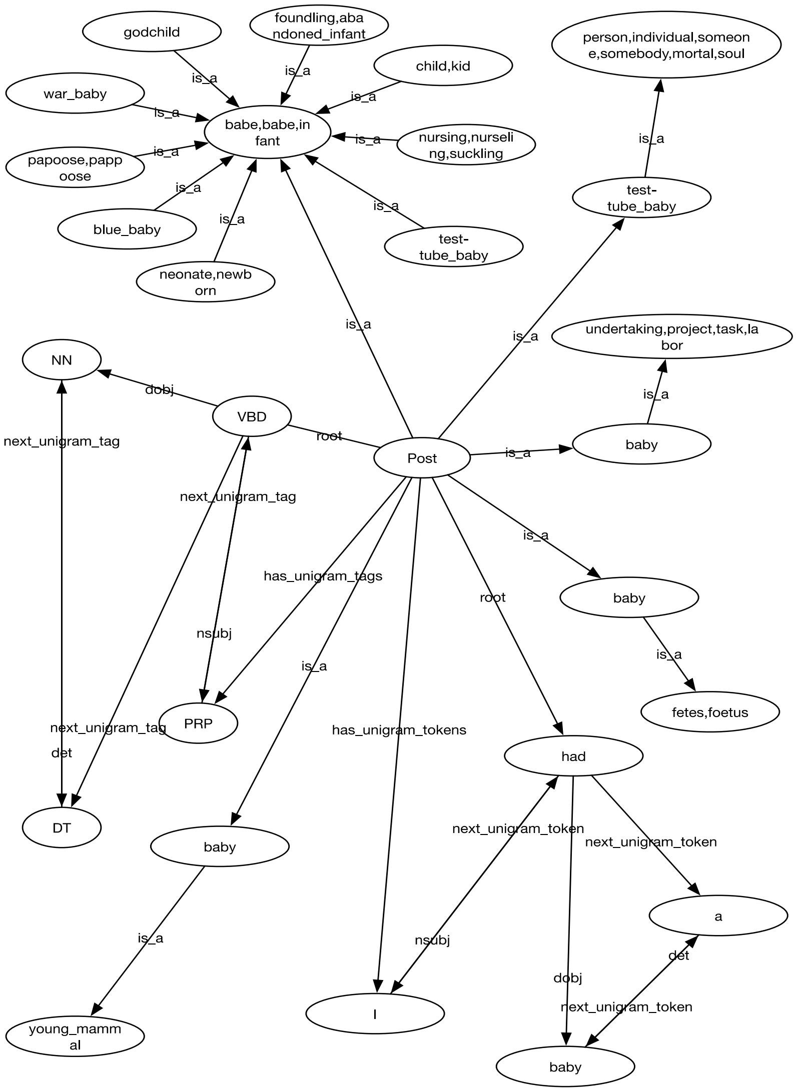

Figure 1. Example Graph of "I had a baby"

4.5 T-Tests

For our T-Tests, we generate 10 samples for each type of event. We create permutations of all possible feature combinations (**N-Grams**, **Entities**, **Topics**, **Syntactic**, **Semantic ConceptNet**, and **Semantic WordNet),** and compare against the best performing baseline combination (mentioned in section 4.4).

Thus, the hypotheses for our t-tests are:

- H0: Graph features do not help improve the f-measure over our baseline.
- H1: By introducing our graph features, we improve our f-measure.

4.6 Feature Reduction

Due to the large number of features extracted from our pattern mining approach, we also consider several feature reduction techniques within Weka. Our approach uses a ranker method, where we filter our attributes via a chosen metric. Initially we remove any feature whose rank is 0, then gradually reduce the number of features over 20 even intervals over the remaining attributes. In order to choose the best suitable attribute metric, we compared the results of several attribute selection algorithms over our Death of a Parent training set: Information Gain, Gain Ratio, Correlation, and Symmetrical Uncertainty. We found that Gain Ratio gave a significant improvement of 13 points.

5. RESULTS

Our results are displayed in Table 1 for all of our events: Getting Married, Death of a Parent, Having Children, Starting School, and Falling in Love. We compare the performance of the best baseline (B) against the best feature combination. We report: F1, change in F1 to baseline (Δ F1), Accuracy (A), Precision (P), Recall (R), P-Value (P-Val). We observe statistical significance when p-val < 0.05.

As can be seen from table 1, in all cases, our graph-based methodology merged with some of the baseline features produces a significant improvement in all measures (F1, Accuracy, Precision, and Recall), across each of our events. Our biggest improvement is in Falling in Love, with a 6.8% percentage point increase, while our smallest is in Having Children, with only 1.1%. However, the improvements over the baseline are all statistically significant (p < 0.01).

Overall our improvements show a 4.54% point increase in F-Measure over our baselines. For each classifier, the best performing baseline combination remains constant (N-Grams, Topics, Entities), which is the full methodology outlined in Li et al. The best performing feature combination is also the same (N-grams, Topics, Entities + WordNet and Syntactic patterns). One explanation for concepts not always appearing in our best classifiers might be that ConceptNet and WordNet have a very similar feature distribution. By adding both, the available features are increased, thus making it more difficult to perform feature selection. Another could be that Concepts may not be effective discriminators.

Table 1. Results Table

Theme	Feature Set	F1	Δ F1	A	P	R	P-Val
Getting Married	Best Baseline: Ent,Ngrm,Tpc	85.9%		86.0%	86.1%	86.0%	
	Con	66.5%	-19.4%	67.8%	67.6%	67.8%	<0.01
	Ent	45.4%	-40.4%	51.4%	63.1%	51.4%	<0.01
	Ngrm	83.7%	-2.2%	84.0%	83.8%	84.0%	<0.01
	Syn	79.0%	-6.9%	79.7%	80.8%	79.7%	<0.01
	Tpc	71.0%	-14.9%	72.3%	71.9%	72.3%	<0.01
	WN	70.6%	-15.3%	71.7%	71.4%	71.7%	<0.01
	Best Graph Only: Con, Syn, Wn	86.3%	0.4%	86.5%	86.5%	86.5%	0.04
	Con,Ent,Ngram,Syn,Tpc,Wn	**90.7%**	**4.8%**	**90.8%**	**90.7%**	**90.8%**	**<0.01**
Death of a Parent	Best Baseline: Ent,Ngrm,Tpc	86.2%		86.2%	86.7%	86.2%	
	Con	71.0%	-15.2%	71.6%	71.5%	71.6%	<0.01
	Ent						
	Ngrm	85.4%	-0.8%	85.5%	86.0%	85.5%	0.02
	Syn	82.2%	-4.0%	82.4%	83.9%	82.4%	<0.01
	Tpc	61.1%	-25.1%	60.4%	67.7%	60.4%	<0.01
	WN	79.4%	-6.7%	79.7%	79.7%	79.7%	<0.01
	Best Graph Only: Con,Syn,WN	87.8%	1.6%	87.9%	88.8%	87.9%	0.06
	Ent,Ngrm,Syn,Tpc,WN	**90.7%**	**4.5%**	**90.7%**	**91.1%**	**90.7%**	**<0.01**
Having Children	Best Baseline: Ent,Ngrm,Tpc	89.8%		89.9%	90.2%	89.9%	
	Con	6.7%	-23.1%	67.7%	67.9%	67.7%	<0.01
	Ent	35.5%	-54.3%	41.4%	59.4%	41.4%	<0.01
	Ngrm	87.9%	-1.9%	88.0%	88.3%	88.0%	<0.01
	Syn	79.7%	-10.1%	80.2%	82.9%	80.2%	<0.01
	Tpc	72.4%	-17.4%	73.1%	72.2%	73.1%	<0.01
	WN	73.8%	-16.0%	74.7%	74.8%	74.7%	<0.01
	Best Graph Only:Con,Syn,WN	86.2%	-3.5%	86.5%	88.0%	86.5%	<0.01
	Ent,Ngrm,Syn,Tpc,WN	**90.9%**	**1.1%**	**91.0%**	**91.7%**	**91.0%**	**<0.01**

Starting School	Best Baseline: Ent,Ngrm,Tpc	87.7%		87.8%	88.3%	87.8%	
	Con	64.9%	-22.8%	66.1%	64.9%	66.1%	<0.01
	Ent	32.1%	-55.6%	37.4%	53.4%	37.4%	<0.01
	Ngrm	85.0%	-2.7%	85.2%	85.2%	85.2%	<0.01
	Syn	86.3%	-1.4%	86.6%	87.9%	86.6%	<0.01
	Tpc	72.4%	-15.3%	72.6%	72.4%	72.6%	<0.01
	WN	66.1%	-21.6%	68.0%	67.0%	68.0%	<0.01
	Best Graph Only:Syn,WN	88.9%	1.2%	89.1%	90.1%	89.1%	0.025
	Ent,Ngrm,Syn,Tpc,WN	**93.0%**	**5.5%**	**93.0%**	**93.5%**	**93.0%**	**<0.01**
Falling in Love	Best Baseline: Ent,Ngrm,Tpc	83.7%		83.9%	84.2%	83.9%	
	Con	68.4%	-15.4%	68.9%	69.4%	68.9%	<0.01
	Ent	29.5%	-54.2%	34.6%	53.3%	34.6%	<0.01
	Ngrm	80.4%	-3.3%	80.6%	81.1%	80.6%	<0.01
	Syn	82.1%	-1.6%	82.3%	84.4%	82.3%	<0.01
	Tpc	62.3%	-21.5%	65.1%	65.3%	65.1%	<0.01
	WN	70.3%	-13.5%	71.3%	71.7%	71.3%	0.103
	Best Graph Only:Con,Syn,WN	86.8%	3.0%	86.9%	87.3%	86.9%	0.03
	Ent,Ngrm,Syn,Tpc,WN	**90.5%**	**6.8%**	**90.6%**	**91.2%**	**90.6%**	**<0.01**

6. DISCUSSION

Our results show that some interesting observations can be made. When combined with our baseline features, our graph F1 scores are consistently and significantly higher. When looking at our top performing features, we found that patterns extracted from ConceptNet and WordNet were regularly at the top, and tended to help discriminate between our negative class, and our two thematic classes. However, when attempting to discriminate theme from event, these types of patterns were less useful as each will tend to use similar languages and concepts. Instead, we found our syntactical patterns helped to discriminate between events and non-events.

While we have successfully identified a number of interesting frequent graph patterns, there is much improvement that can still be made on this. Our choice of algorithm does not scale well, thus we have had to perform some optimization steps that might limit our results. Newer state of the art algorithms (e.g., (Pan, et al., 2015)) also consider the feature selection problem as being part of the pattern mining discovery phase, reducing memory and time cost. An alternative might be to consider using a more distributed pattern mining approach, such as outlined in (Bhuiyan & Hasan, 2014). In addition to this, our assumption that two concepts that are close together in ConceptNet are similar is not always valid. For example, the relationship “IsAntonym” in ConceptNet is the opposite of similar in terms of linking two concepts together. When not taking this into account, concepts such as “life” and “death” might be considered semantically similar. Thus a better approach would be to either filter out all negating relationships like this one, or to add a weighting system where a heuristic function could possibly be used to calculate how close two concepts are together in the graph (Rada, et al., 1989). Future work can also consider expanding our semantic graphs into other networks, such as DBpedia. We could also consider other types of semantic features such as semantic frames using FrameNet (Baker, et al., 1998).

In this work, we focused on the identification of only five of the most common life events. A natural next step is to expand to the full set of life events listed in (Janssen & Rubin, 2011). This would help test the consistency of our results across other types of event. A broader goal would be to consider the detection of life events in general. In addition to this, we construct a classifier for each individual life event. As Li et al (Li, et al., 2014) point out in their work, a single large multi classifier is a more complex task compared to individual theme detection, due to the overlap of n-grams in certain topics. Given our observation of how our semantic patterns work, it would be interesting to see if they can help boost the results in such a classifier.

We opted to serialize our frequent patterns into a key/map pair so that we can represent our training set as a vector of binary values. However, there exist a number of alternative classification algorithms that are designed specifically for graphs; some reliant on frequent patterns themselves, others taking into account the structure of the graph. In terms of our frequent patterns, we also see a number of weak infrequent ones. However, rather than representing these patterns as independent of each other, we could consider a methodology such as COM (Jin, et al., 2009), where co-occurrence of disconnected low frequency patterns is considered instead, converting them into higher frequency ones.

7. CONCLUSION

We have presented a new approach for the problem of life event detection that focusses on five major life events identified by psychologists: Getting Married, Having Children, Death of a Parent, Starting School, and Falling in Love. Our approach expands a Twitter post into both a syntactic and semantic graph. This network is then mined for frequent sub patterns that can be used as features within a classifier, filtered via their gain ratio. Our results showed consistent and significant improvement over baselines from other work, that consider only features extracted directly from the post (i.e., no graphs). We showed that our graph-based approach achieves significantly better accuracy and F1 scores than our selected baseline.

REFERENCES

Baker, C. F., Fillmore, C. J. & Lowe, J. B., 1998. *The berkeley framenet project.* s.l., s.n.

Bhuiyan, M. A. & Hasan, M. A., 2014. *FSM-H: Frequent Subgraph Mining Algorithm in Hadoop.* s.l., s.n., pp. 9-16.

Cao, B. et al., 2015. Mining brain networks using multiple side views for neurological disorder identification. *arXiv preprint arXiv:1508.04554.*

Chen, D. & Manning, C. D., 2014. *A Fast and Accurate Dependency Parser using Neural Networks..* s.l., s.n., pp. 740-750.

Choudhury, S. & Alani, H., 2014. *Personal life event detection from social media.* s.l., CEUR.

Covington, M. A., 2001. *A fundamental algorithm for dependency parsing.* Athens, s.n.

Di Eugenio, B., Green, N. & Subba, R., 2013. *Detecting life events in feeds from twitter.* s.l., s.n.

Dickinson, T. et al., 2015. *Identifying Prominent Life Events on Twitter.* s.l., s.n.

Holder, L. B., Cook, D. J., Djoko, S. & others, 1994. *Substucture Discovery in the SUBDUE System..* s.l., s.n., pp. 169-180.

Janssen, S. M. J. & Rubin, D. C., 2011. Age Effects in Cultural Life Scripts. *Applied Cognitive Psychology,* 25(2), pp. 291-298.

Jin, N., Young, C. & Wang, W., 2009. Graph classification based on pattern co-occurrence. *Proc. Conf. on Info. and Knowledge Man. - CIKM '09.*

Kong, L. et al., 2014. *A dependency parser for tweets.* s.l., s.n.

Li, J., Ritter, A., Cardie, C. & Hovy, E., 2014. *Major Life Event Extraction from Twitter based on Congratulations/Condolences Speech Acts.* Doha, s.n.

Liu, X., Troncy, R. & Huet, B., 2011. *Using social media to identify events.* s.l., s.n.

Ohara, K. et al., 2008. *Pruning Strategies Based on the Upper Bound of Information Gain for Discriminative Subgraph Mining..* Vietn, s.n.

Owoputi, O. et al., 2013. *Improved part-of-speech tagging for online conversational text with word clusters.* s.l., s.n.

Pan, S. et al., 2015. Finding the best not the most: regularized loss minimization subgraph selection for graph classification. *Pattern Recognition,* Volume 48, pp. 3783-3796.

Parameswaran, A., Rajaraman, A. & Garcia-Molina, H., 2010. Towards the web of concepts: Extracting concepts from large datasets. *Proceedings of the Very Large Data Bases Conference (VLDB)* , 3((1-2)).

Philippsen, M., Worlein, M., Dreweke, A. & Werth, T., 2011. Parsemis: the parallel and sequential mining suite. *Available at www2. informatik. uni-erlangen. de/EN/research/ParSeMiS.*

Rada, R., Mili, H., Bicknell, E. & Blettner, M., 1989. *Development and application of a metric on semantic nets.* s.l., s.n., pp. 17-30.

Resnik, P., 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research,* Volume 11, pp. 95-130.

Sakaki, T., Okazaki, M. & Matsuo, Y., 2010. *Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors.* Raleigh, North Carolina USA, s.n.

Sayyadi, H., Hurst, M. & Maykov, A., 2009. *Event detection and tracking in social streams.* San Jose, {CA}, s.n.

Wayne, C. L., 2000. *Topic detection and tracking in English and Chinese.* NY, s.n.

Yan, X. & Han, J., 2002. *gspan: Graph-based substructure pattern mining.* s.l., s.n., pp. 721-724.

Yan, X. & Han, J., 2003. *CloseGraph: mining closed frequent graph patterns.* Washington, s.n.