DETERMINING WHETHER AND WHEN PEOPLE PARTICIPATE
IN THE EVENTS THEY TWEET ABOUT

Krishna Chaitanya Sanagavarapu, B.Tech

Thesis Prepared for the Degree of

MASTER OF SCIENCE

UNIVERSITY OF NORTH TEXAS

May 2017

APPROVED:

Eduardo Blanco, Major Professor
Xiaohui Yuan, Committee Member
Yan Huang, Committee Member
Barrett Bryant, Chair of the Department of
Computer Science and Engineering
Coastas Tsatsoulis, Dean of the College of
Engineering
Victor Prybutok, Vice Provost of the
Toulouse Graduate School
Sanagavarapu, Krishna Chaitanya. *Determining Whether and When People Participate in the Events They Tweet About*. Master of Science (Computer Science), May 2017, 30 pp., 13 tables, 8 figures, 19 numbered references.

This work describes an approach to determine whether people participate in the events they tweet about. Specifically, we determine whether people are participants in events with respect to the tweet timestamp. We target all events expressed by verbs in tweets, including past, present and events that may occur in future. We define event participant as people directly involved in an event regardless of whether they are the agent, recipient or play another role. We present an annotation effort, guidelines and quality analysis with 1,096 event mentions. We discuss the label distributions and event behavior in the annotated corpus. We also explain several features used and a standard supervised machine learning approach to automatically determine if and when the author is a participant of the event in the tweet. We discuss trends in the results obtained and devise important conclusions.
Copyright 2017
by
Krishna Chaitanya Sanagavarapu
ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my major professor, Dr. Eduardo Blanco, who has supported me in this thesis journey through his patience and knowledgeable inputs. I attribute my Master’s degree to his encouragement and support.

I would also like to express my gratefulness to Dr. Xiaouhi Yuan and Dr. Yan Huang for their support in providing me with useful inputs to my thesis. I would love to extend my thanks to my lab mates for their great support and help in my thesis. My special thanks to my friends who gave their valuable inputs in designing my thesis providing their valuable feedback. I would also like to extend my gratitude to Dr Barret Bryant, Stephanie Deacon, Department of Computer Science and UNT for providing me useful resources and supporting me that made my work easy.

Certainly, this thesis would not have been possible without the unconditional love and support of my parents and relatives. I wish to dedicate this thesis to my family.
# TABLE OF CONTENTS

ACKNOWLEDGMENTS iii  
LIST OF TABLES vi  
LIST OF FIGURES vii  

CHAPTER 1 INTRODUCTION 1  
1.1. Problem Description 1  
1.2. Contributions 2  
1.3. Thesis Organization 2  

CHAPTER 2 PREVIOUS WORK 3  
2.1. Event Extraction from Twitter 3  
2.2. Twitter Tools 4  

CHAPTER 3 CORPUS CREATION 7  
3.1. Gathering a Corpus of Tweets 7  
3.2. Pre-processing Tweets 7  
3.3. Annotation Process 9  
3.4. Annotation Guidelines 10  
3.5. Annotation Quality 12  
3.6. Annotation Examples 14  

CHAPTER 4 CORPUS ANALYSIS 16  
4.1. Label Distribution 16  
4.2. Event Analysis 17  

CHAPTER 5 EXPERIMENTS AND RESULTS 19  
5.1. Experimental Setup 19  
5.2. Features Used 19
5.3. Experimental Results

5.3.1. Results with fine grained labels

5.3.2. Results with coarse grained labels

CHAPTER 6 CONCLUSIONS

BIBLIOGRAPHY
# LIST OF TABLES

| Table | Description                                                                 | Page |
|-------|-----------------------------------------------------------------------------|------|
| 3.1   | Counts of tweets and events gathered for annotating.                        | 7    |
| 3.2   | Inter-annotator Kappa agreement. The table shows Fleiss Kappa, observed agreement and expected agreement scores for each temporal anchor and the overall scores. | 12   |
| 3.3   | Percent disagreements. The table shows the percent of disagreements for 10% of selected events in the corpus for each temporal anchor and overall. | 13   |
| 5.1   | Features to determine whether the author of a tweet is a participant in the events he tweets about. | 20   |
| 5.2   | Feature values for an example tweet.                                        | 21   |
| 5.3   | Results with most frequent baseline per temporal span using fine grained labels | 22   |
| 5.4   | Results with event features per temporal span using fine grained labels      | 22   |
| 5.5   | Results with event and situation entities features per temporal span using fine grained labels | 23   |
| 5.6   | Results with event, situation entities and context features per temporal span using fine grained labels | 23   |
| 5.7   | Results with most frequent baseline per temporal span using coarse grained labels | 24   |
| 5.8   | Results with event features per temporal span using fine grained labels      | 24   |
| 5.9   | Results with event and situation entities features per temporal span using coarse grained labels | 25   |
| 5.10  | Results with event, situation entities and context features per temporal span using coarse grained labels | 25   |
# LIST OF FIGURES

| Figure | Description                                                                                   | Page |
|--------|-----------------------------------------------------------------------------------------------|------|
| 2.1    | Output of TWICAL on example tweet. We obtained POS tags, event and chunks from the tool.       | 5    |
| 2.2    | Output of TweeboParser on example tweet.                                                       | 6    |
| 2.3    | Combining the output from TWICAL and TweeboParser.                                             | 6    |
| 3.1    | Annotation interface provided to the annotators.                                               | 9    |
| 4.1    | Distribution of labels per temporal span. Percentages are shown if they are greater or equal than 2%, e.g., percentages for unk label are not shown because they range between 0.91% and 1.91%. | 16   |
| 4.2    | Overall distribution of labels per part-of-speech tags of the event.                           | 17   |
| 4.3    | Overall distribution of labels for top 5 frequent events.                                      | 18   |
| 5.1    | Dependency graph obtained from tweebo parser for the example tweet                            | 21   |
CHAPTER 1

INTRODUCTION

Twitter has quickly become one of the most popular social media sites. It has 313 million monthly active users, and 500 million tweets are published daily. People tweet about many things, they tweet to express opinions and ideas e.g., #HarryPotter is the greatest fantasy film!! about events e.g., Super excited for the concert!! to advertise something e.g., Style your spring closet with #Express 20% off everything!!!, quotes, humor or articles e.g., Whatever doesn’t kill you makes u stranger!! #jokerQuotes, to express wishes, condolences etc. e.g., All the best for finals @john.

Natural Language Processing systems can leverage the wide range of information that is present in tweets to make way for applications like recommender systems that recommend nearby events in which the user might be interested in, targeted online advertising, friend recommendations by matching users with similar opinions, ideas or who went to the same school etc.

1.1. Problem Description

People tweet about events (underlining indicates events of interest below): in which they have participated in the past ( e.g., Loved my #SpringBreak trip), are participating in the present ( e.g., having fun at #ATPTours @john missing you here buddy!) or might participate in future ( e.g., Hope to get tickets for the #OneDirection concert this time #fingerscrossed), they are promoting ( e.g., every 1 pls share nd support the 12k run #RunForFitness), in which their friends, relatives or acquaintances have participated ( e.g., My nephew finally graduated! #congrats @Joe) or world and local events in general ( e.g., Saddened by the horrors of the #katrina hurricane).

People tweet not only about events in which they participate, but also events in which they do not participate but are somehow relevant ( e.g., John Doe may tweet about his nephew graduating from college). More specifically, people can participate in the events they tweet about prior to tweeting ( e.g., When I come back to London, I realize how much I miss living here), while tweeting ( e.g., Nope. Not yet. Still in my car, enjoying traffic), or after tweeting ( e.g., Can’t wait to fly home this summer). In the third example, it is not guaranteed that fly will occur, so one can
only say that the author will probably participate in fly.

In this work, we determine if the author of a tweet participates in the events that he tweets about. We also determine if the author is a participant in that event before tweeting, while tweeting and after tweeting. We define event participants as people directly involved in an event, regardless of whether they are agent, recipient or play another role.

1.2. Contributions

The major contributions of this work are:

1. We annotate a corpus of 826 tweets with 1096 event mentions with the following information:
   - whether or not the author of the tweet is a participant in the event mention
   - when is the author a participant of the event with respect to tweet timestamp

2. We study the annotated corpus and present:
   - analysis of the label distribution in the corpus
   - behavior of several events in the corpus

3. We detail the supervised learning experiments performed to automatically determine if and when the author of a tweet participates in the event he tweets about

1.3. Thesis Organization

This thesis is organized as follows:

1. In chapter 2 we detail the work related to event extraction and introduce the novelties in this work. We also explain the tools used in our work.

2. In chapter 3 we present how we gather an initial corpus of tweets and pre-process them for annotating. We describe the annotation process and guidelines followed while annotating. We also discuss the quality of annotations in this chapter.

3. In chapter 4 we examine the created corpus and present the analysis of the label distribution and behavior of events in the corpus

4. In chapter 5 we present experimental setup, detail the features used for our experiments and discuss the results obtained with standard supervised learning.

5. We present important conclusions in chapter 6
CHAPTER 2

PREVIOUS WORK

In this chapter we discuss the previous efforts on event extraction (section 2.1). We also explain how our work is unique from the previous efforts. In section 2.2, we list certain tools developed for Twitter that we use in our work.

2.1. Event Extraction from Twitter

Most previous efforts in detecting events from Twitter focus on events of general importance (death of a celebrity, natural disasters) or major life events of individuals (John Doe getting married, having a baby, being promoted).

Dickinson et. al. [3] make an effort to identify top five life events in Twitter: having children, beginning school, marriage, parent’s death and falling in love. They demonstrate how user content and semantic features can be applied in the identification of prominent life events. They also make available 2,241 manually annotated tweets for life event identification. Their best supervised classifier J48 uses semantic and n-gram features and performs with an F1 score of 0.73.

Barbara et. al. [2] determine if tweets about employment or marriage affects the tweeter or another entity in the past or present or if it is a general statement. They study uni-gram models and also train machine learning classifiers with parts of speech tag, semantic and context features. They could achieve an accuracy of 74% with marriage and 90% with employment events.

Major life event extraction from Twitter based on congratulations/condolences speech acts [9] demonstrates extraction of fine-grained description of user's life events based on their published tweets. Extracting major life events consists of pinpointing significant events from mundane events (having lunch, exercising) and determining whether significant events are relevant to Twitter users (Why doesn’t John marry Mary already? [not relevant to the author] vs. Today is my 10-year marriage anniversary [relevant to author]. Their pipeline system identifies major life event category for input tweets, determines if the speaker is directly involved in the life event and extracts the property of the event. They construct a gold standard life event dataset with 900 positive tweets.
They present supervised and bootstrapping approaches. Their best system achieves a precision of 0.65.

Extracting events of general importance includes extracting entities involved, date and location, and classifying events into classes such as trail, product launch or death. Given a raw stream of tweets TWICAL [16] system extracts named entities in association with event phrases and unambiguous dates which are involved in significant events. First, the tweets are POS tagged, then named entities and event phrases are extracted, temporal expressions resolved, and the extracted events are categorized into types. Finally, the strength of association between each named entity and date is measured based on the number of tweets they co-occur in, in order to determine whether an event is significant.

Exploiting redundancy in tweets to extract events is common. Bayesian modeling approach which is able to directly extract event-related keywords from tweets without supervised learning proposed in [19] outperforms TWICAL in open event extraction.

Extracting Spatio-temporal information involves determining when and where tweets originate from [1] which suggest an alternative methodology for event detection using space-time scan statistics (STSS). This technique looks for clusters within the dataset across both space and time, regardless of tweet content. It is expected that clusters of tweets will emerge during spatio-temporally relevant events, as people will tweet more than expected in order to describe the event and spread information.

Unlike these previous efforts, our work:

- determines whether people participate in the events they tweet about
- specifies when with respect to tweet timestamps
- targets past events, ongoing events and events likely to occur in the future
- targets all events regardless of importance

2.2. Twitter Tools

In this section, we discuss the tools developed specifically for twitter data from which we extract events to annotate and features while performing learning experiments. We also give example output of each tool and explain how we utilized them.
FIGURE 2.1. Output of TWICAL on example tweet. We obtained POS tags, event and chunks from the tool.

a) T-POS: Parts of speech tagger trained on in-domain twitter data [15]. T-POS:
   - uses conditional random fields and employs brown clusters, POS dictionaries, spelling and contextual features.
   - uses Penn TreeBank tagset as well as the following additional tags for twitter specific phenomena: re-tweets, hash-tags, usernames and URLs.
   - when trained on 102K tokens from twitter, IRC chat data and Penn TreeBank obtains an accuracy of 0.88 and a 41% error reduction over Stanford POS tagger.

b) We use the event tagger of TWICAL[16] a system that extracts the 4-tuple representation of events which includes a named entity, event phrase, calendar date and event type. The event tagger:
   - is trained on 1000 tweets (19,484 tokens) annotated with event phrases.
   - uses conditional random fields with contextual, dictionary, orthographic features and features from twitter tuned parts of speech tagger.
   - Obtains an F-score of 0.64.
   - Figure 2.1 shows the output of TWICAL system from which we utilized the events information for tweets.
   - For a given tweet the output contains tokenized and tagged words separated by forward slash.

c) TweeboParser, a dependency parser for twitter[6]:
   - TweeboParser predicts the syntactic structure represented by unlabeled dependencies for a
given tweet.

- It produces a multi-rooted graph over a tweet and excludes hash-tags, emoticons and URLs that do not have a syntactic function, from the parse tree.
- TweeboParser modifies TurboParser [11] and adds additional Brown clusters and Penn TreeBank features to it.
- TweeboParser achieves over 80% unlabeled attachment score.
- We use TweeboParser to obtain dependencies which are used to extract certain features in our learning experiments. Figure 2.2 portrays sample output of the tool. Column 5 of the output gives the token indices of the token to which the current token has a dependency. This is the unlabeled dependency graph.

Figure 2.3 is the combined output that we generated by merging the outputs from TWICAL and TweeboParser. We used this combined output to extract features for our learning experiments. We utilize the POS tags (column 8), event information (column 10) and dependencies (column 5) from the output

\[
\begin{array}{cccccccccc}
\text{#HarryPotter} & ^ & ^ & _ & 2 & _ & 0 & \text{NNS} & \text{B-NP} & 0 \\
\text{Marathon} & _ & \text{N} & \text{N} & _ & 0 & _ & 0 & \text{NNP} & \text{I-NP} & 0 \\
\text{while} & _ & \text{P} & \text{P} & _ & 2 & _ & 0 & \text{IN} & \text{B-SBAR} & 0 \\
\text{I} & _ & \text{O} & \text{O} & _ & 5 & _ & 0 & \text{PRP} & \text{B-NP} & 0 \\
\text{write} & _ & \text{V} & \text{V} & _ & 3 & _ & 0 & \text{VB} & \text{B-VP} & \text{B-EVENT} \\
\text{my} & _ & \text{D} & \text{D} & _ & 7 & _ & 0 & \text{PRP}\$ & \text{B-NP} & 0 \\
\text{papers} & _ & \text{N} & \text{N} & _ & 5 & _ & 0 & \text{NNS} & \text{I-NP} & 0 \\
. & _ & , & , & _ & -1 & _ & 0 & . & 0 & 0 \\
\end{array}
\]

**Figure 2.2.** Output of TweeboParser on example tweet.

\[
\begin{array}{cccccccccc}
\text{#HarryPotter} & ^ & ^ & _ & 2 & _ & 0 & \text{NNP} & \text{B-NP} & 0 \\
\text{Marathon} & _ & \text{N} & \text{N} & _ & 0 & _ & 0 & \text{NNP} & \text{I-NP} & 0 \\
\text{while} & _ & \text{P} & \text{P} & _ & 2 & _ & 0 & \text{IN} & \text{B-SBAR} & 0 \\
\text{I} & _ & \text{O} & \text{O} & _ & 5 & _ & 0 & \text{PRP} & \text{B-NP} & 0 \\
\text{write} & _ & \text{V} & \text{V} & _ & 3 & _ & 0 & \text{VB} & \text{B-VP} & \text{B-EVENT} \\
\text{my} & _ & \text{D} & \text{D} & _ & 7 & _ & 0 & \text{PRP}\$ & \text{B-NP} & 0 \\
\text{papers} & _ & \text{N} & \text{N} & _ & 5 & _ & 0 & \text{NNS} & \text{I-NP} & 0 \\
. & _ & , & , & _ & -1 & _ & 0 & . & 0 & 0 \\
\end{array}
\]

**Figure 2.3.** Combining the output from TWICAL and TweeboParser.
CHAPTER 3

CORPUS CREATION

In this chapter, we present how we gathered and annotated a corpus of event mentions in tweets with author participation information. In section 3.1 we discuss the sources from which the initial corpus of tweets are obtained. We describe how we pre-processed the tweets to obtain the final set of event mentions in tweets for annotating in section 3.2. We explain several guidelines followed while annotating the event mentions for author participation in section 3.4. We discuss the quality of the annotations in section 3.5 and present annotation examples in section 3.6.

3.1. Gathering a Corpus of Tweets

In this work, we determine if and when the author of a tweet participates in the event he tweets about. We hence collected tweets that tend to have a good number of events. Unlike previous works that consider major life events [9] or real time events [17], we target all events regardless of importance. We gathered tweets from the following sources:

1. Tweets released by Alan Ritter [15][16] that were used to train and evaluate a named entity recognizer and event extraction tool for twitter. A total of 1993 tweets is collected from this work.
2. Tweets used for training TweeboParser [6], a dependency parser for twitter are also gathered. This data consisted of 703 tweets.
3. Another major source that we considered is ArkTweetNLP [13] comprising of 2321 tweets that are used to train parts of speech tagger for twitter.

A total of 5017 tweets were obtained from the above-mentioned sources.

3.2. Pre-processing Tweets

| Filter | Tweets | Events |
|--------|--------|--------|
| Initial | 5017   | 7008   |
| Filter 1: tweets that contain the words I, me or we | 3267   | 4235   |
| Filter 1 + events with POS tag VBP, VBG and VBN | 826    | 1096   |

TABLE 3.1. Counts of tweets and events gathered for annotating.
We extracted events from the 5017 gathered tweets by using the event extraction tool for twitter [16]. A total of 7008 events were extracted from these tweets. We then performed pilot annotations on 10% of these events. For each event, we asked the annotators to annotate if and when the author of the tweet participated in the event. From this effort, we realized that determining if their author of the tweet participates in an event is not valid for all events. In the tweet, “Had nightmare where I was trying to get to a final exam but somehow got trapped in a game show” asking if the author participated in the event final is not valid.

We decided to add filters in order to extract tweets with events in which the author would likely participate.

We found it intuitive to consider the tweets that contain the pronouns I, me or we. The author of a tweet is likely to participate in the event mentioned in the tweet whenever he/she tweets saying I, me or we. In tweets like “@abzmedic I’m missing autumn because of work. :( I must get into the forests tomorrow!” presence of the pronoun I indicate possible participation of the author in the event missing. However, the presence of I, me or we does not guarantee the participation of the author in the event. In the tweet “We missed my nephews graduation # sad” the author does not participate in the event graduation.

We obtained a total of 3267 tweets with 4235 events after applying this filter.

We also analyze the parts of speech tags of the events and found that the event recognizer makes more mistakes with nouns than verbs. In the tweet ”I had a substitute today that looked like @thedillon from the side of his face. Haha” the noun substitute is recognized as an event but in fact it is not.

Hence we only considered those events with parts of speech tags

1. VBP- verb, non-3rd person singular present
2. VBG- verb, gerund or present participle
3. VBN- verb, past participle

We eliminate those events with parts of speech tag VBD (verb, past tense), VBZ (verb, 3rd person singular present) and VB (verb, base form) since the participation of the author in these events were minimal. In the tweet ”before the season even starts I will not respond to any new
found Miami heat fans” the event *starts* with parts of speech tag VBZ is recognized as an event but the author does not participate in the event *starts*. In the tweet ”*Bored at work. Looking like another 8 hour day. Talk to me*” the event *Talk* with parts of speech tag VB is recognized as an event but the author does not participate in the event *Talk*. Similarly, in the tweet ”*@DeeNeeCole yea I was wondering if after they banned the Indian we got a new one*” the event *banned* with parts of speech tag VBD is recognized as an event but the author does not participate in the event *banned*.

Table 3.1 summarizes the counts of tweets and events after applying each filter to the initial dataset. We obtain 826 tweets after applying all the filters(Table 3.1 row 3). This final set of tweets comprised of 1096 events that we annotated.

### 3.3. Annotation Process

The goal of this work is to determine if and when the author of the tweet participates in the event mentioned in the tweet. After pilot annotations, we found it feasible to ask the question: ”Is the author of the tweet a participant in the event?” for the following temporal spans:

1. Over 24 hours before tweeting (≥24 hrs before)
2. Within 24 hours before tweeting (<24 hrs before)
3. During tweeting (tweet timestamp)
4. Within 24 hours after tweeting (<24 hrs after)
5. Over 24 hours after tweeting (≥24 hrs after)
We allowed six answers for each temporal span, partially inspired by previous work on factuality [18]:

1. Certainly Yes/No (cYes/cNo): I am certain that the author is (or is not) a participant in the event
2. Probably Yes/No (pYes/pNo): It is probably the case that the author is (or is not) a participant in the event
3. Unknown (unk): The question is intelligible, but none of the four labels above would be correct
4. Invalid (inv): The event at hand is not a valid event

Figure 3.1 depicts the interface that we provided the annotators in order to annotate each event with their corresponding labels for all the five temporal spans.

3.4. Annotation Guidelines

We gave the following guidelines for the annotators to answer the question described in section 3.3:

1. In a tweet with a negated event, annotate if the author is involved in the negated event.
   a) I am not running tomorrow.
      • ≥24h Before: unk, <24h Before: unk, event timestamp: cNo, <24h After: cNo, ≥24h After: unk.

2. In tweets with events that could happen in future, the labels will be either pYes/pNo/unk for after.
   a) I plan to attend the concert tomorrow
      • ≥24h Before: cNo, <24h Before: cNo, event timestamp: cNo, <24h After: pYes, ≥24h After: unk.

3. In tweets where author questions for the occurrence of an event where he is involved/not involved in the event, choose the label pYes if the event is not negated.
   a) :( RT @themaine Who is coming to the show tomorrow in Hawaii?
      • ≥24h Before: cNo, <24h Before: cNo, event timestamp: cNo, <24h After: pYes, ≥24h After: pYes.
4. For re-tweets of news or other events being reported the label is cNo for all three temporal anchors
   a) @ChefGuyFieri Pls RT - NIGHT golf tournament 9/25 to benefit 3 year old Penny of Mesa, AZ who had heart transplant. www.flancers.com 4 info
      • ≥24h Before: cNo, <24h Before: cNo, event timestamp: cNo, <24h After: cNo, ≥24h After: cNo.

5. Authors are not involved in events that happen due to someone/something external
   a) @BornThisWayBaby thankk yhuu # monsterlove u just made my day :D
      • ≥24h Before: cNo, <24h Before: cNo, event timestamp: cNo, <24h After: cNo, ≥24h After: cNo.
   b) @ConorMc GaGa played here a few weeks back. She spoke out against SB1070 during the show. God bless her for that.
      • ≥24h Before: cNo, <24h Before: cNo, event timestamp: cNo, <24h After: cNo, ≥24h After: cNo.

6. For events that are states, once the author reached that state the three temporal anchors are annotated as cYes/cNo depending on the meaning
   a) @davidcushman David, I just realized the other day you’re no longer at Brando ? How come ? Doing your own thing ?! ( ps yeah it sucks ! ) hre, the author realized something so he will be in that state in future also, he cannot un-realize
      • ≥24h Before: pNo, <24h Before: cYes, event timestamp: cYes, <24h After: cYes, ≥24h After: cYes.
   b) @DEVEY2G no I liked the one u had had that time but I’m sure I cood make one wit my bro possibly here, (the author liked something so he will be liking it in future as well)
      • ≥24h Before: cYes, <24h Before: cYes, event timestamp: cYes, <24h After: cYes, ≥24h After: cYes.

7. For events that happen only when some other condition is met, the involvement of author is decided assuming the condition is met.
   a) @ethernat You know me, if I can be kind and guilt you at the same time, well, my job is
done :)

- ≥24h Before: unk, <24h Before: unk, event timestamp: cNo, <24h After: cYes, ≥24h After: cYes.

8. For the events whose occurrence is periodic, consider that particular instance of the occurrence in the given tweet. Eg. Fathers’ Birthday.

9. If an event is originally a verb and is not valid (say not a verb) in the context of the sentence, annotate as inv.

10. If the author is tentative about the participation in a future event in his tweet, we consider him as a participant (leaving that task to factuality).

    a) @jackpittgregor Tomorrow is a rerun. It’s before their time. I don’t know about Nate or Ford. I forgot what year it’s from. I am thinking ...

        - ≥24h Before: cNo, <24h Before: cNo, event timestamp: cNo, <24h After: pYes, ≥24h After: pYes.

11. If the event is misspelled, then label it as inv for all temporal tags.

    a) I’m in a bar and checking my twit acc so i guess its time to go home.

        - ≥24h Before: inv, <24h Before: inv, event timestamp: inv, <24h After: inv, ≥24h After: inv.

3.5. Annotation Quality

|                      | Before                      | Tweet timestamp | After                      | Overall                      |
|----------------------|-----------------------------|-----------------|----------------------------|------------------------------|
|                      | ≥= 24 hrs                  | <24 hrs         | <24 hrs                    | ≥= 24 hrs                    |
| Fleiss Kappa         | 0.61                        | 0.78            | 0.74                       | 0.70 0.62 0.70               |
| Observed Agreement   | 0.81                        | 0.87            | 0.88                       | 0.81 0.77 0.83               |
| Expected agreement   | 0.52                        | 0.43            | 0.56                       | 0.38 0.40 0.45               |

Table 3.2. Inter-annotator Kappa agreement. The table shows Fleiss Kappa, observed agreement and expected agreement scores for each temporal anchor and the overall scores.

We calculate Kappa scores and percentage disagreements on 10% of events to assess the quality of the annotations. The remaining events are annotated once.

Table 3.2 presents the Kappa values for each temporal span. The overall Kappa value is 0.70. Kappa agreement between 0.60 and 0.80 is considered substantial agreement [8].
The agreement for before \( \geq 24 \text{ hrs} \) (0.61) is less compared to before \(< 24 \text{ hrs}\) (0.78). Also, the agreement for after \( \geq 24 \text{ hrs}\) (0.62) is less compared to after \(< 24 \text{ hrs}\) (0.70). From this trend, we conclude that it is difficult to determine whether the author participates in the event for over 24 hrs before and after tweeting.

|                      | Before \( \geq 24 \text{ hrs}\) | Before \(< 24 \text{ hrs}\) | Tweet timestamp | After \(< 24 \text{ hrs}\) | After \( \geq 24 \text{ hrs}\) | Overall |
|----------------------|----------------------------------|----------------------------|-------------------|---------------------------|---------------------------|---------|
| % total disagreement | 18.10                            | 12.07                      | 11.21             | 17.24                     | 22.41                     | 16.21   |
| % disagreement between Yes and No | 8.62                          | 4.31                      | 6.90              | 4.31                      | 6.90                      | 6.21    |
| % disagreement between CYes and PYes or Cno and PNo | 1.72                          | 1.72                      | 0.00              | 7.76                      | 7.76                      | 3.79    |
| % other disagreement | 7.76                            | 6.03                      | 4.31              | 5.17                      | 7.76                      | 6.21    |

**Table 3.3.** Percent disagreements. The table shows the percent of disagreements for 10% of selected events in the corpus for each temporal anchor and overall.

In Table 3.3 we present:

1. the total % disagreement (row1): percentage of instances when the two annotators disagree on any label with respect to all instances
2. % disagreement between yes and no (row 2): percent of instances when one annotator choose yes\((cYes/pYes)\) and the other annotator choose no\((cNo/pNo)\) among all disagreed instances
3. % disagreement between \(cYes\) and \(pYes\) or \(cNo\) and \(pNo\) (row 3): percent instances where one of the annotator choose probably\((yes/no)\) and the other choose certainly \((yes/no)\).
4. % other disagreements (row 4): percent of instances where the annotators disagree between the following labels:
   - \(inv\) and \(unk\)
   - \(yes\,(cYes/pYes)\) and \(inv\)
   - \(no\,(cNo/pNo)\) and \(inv\)
   - \(yes\,(cYes/pYes)\) and \(unk\)
   - \(no\,(cNo/pNo)\) and \(unk\)

The % of total disagreement is less for tweet timestamp compared to other temporal spans. Also, the % disagreement between certainly\((yes/no)\) and probably\((yes/no)\) is 0 in case of tweet timestamp. From this observation, we conclude that determining if the author is a participant of
the given event while tweeting is easy.

The disagreement between certainly (yes/no) and probably (yes/no) is higher compared to the disagreement between yes (cYes/pYes) and no (cNo/pNo) for the temporal spans ≥24 hrs and <24 hrs after. Whereas, the opposite is true for the temporal spans ≥24 hrs and <24 hrs before. The % total disagreement for over 24 hrs is greater than for within 24 hrs for both before (18.10 vs. 12.07) and after (22.41 vs. 17.24). Hence, answering for over 24hrs is difficult. The % total disagreement is the highest (22.41) for ≥24 hrs after making it the most difficult question to answer.

3.6. Annotation Examples

This section presents real annotation examples. In the Tweet *Im so addicted to Twitter now that I can tweet all the time. Not good*, the annotators choose pYes for before ≥24 hrs, cYes for <24 hrs before, cYes for during tweet time stamp, cYes for <24 hrs after and pYes for ≥24 hrs after. Annotators understood that the author has certainly been addicted to Twitter for 24 hours before and after tweeting, and probably longer.

In the Tweet *I wonder if you realize we were talking about YOU*, the annotators choose pNo for before ≥24 hrs, cYes for <24 hrs before, cNo for during tweet time stamp, cNo for <24 hrs after and cNo for ≥24 hrs after. The author of Tweet was clearly talking about YOU before but not after tweeting and not during tweeting; annotators indicated that talking most likely occurred within 24 hours before tweeting.

In the Tweet *I’m selling my snorkle, text me for details*, the annotators choose cNo for before ≥24 hrs, pYes for <24 hrs before, cYes for during tweet time stamp, cYes for <24 hrs after and pYes for ≥24 hrs after. The annotations in Tweet indicate that the author was certainly a participant of selling when he tweeted and within 24 hours after tweeting, and probably also within 24 hours before and over 24 hours after. They indicated the author is not a participant of selling ≥24 hrs before tweeting.

In the Tweet *@Genuine Will I be seeing you at #typeamom next week?*, the annotators choose cNo for before ≥24 hrs, cNo for <24 hrs before, cNo for during tweet time stamp, cNo for ≥24 hrs after.
for <24 hrs after and pYes for ≥24 hrs after. Event seeing in the Tweet may occur next week, and annotations capture this information (all cNo except 24h after, which is pYes).

In the Tweet Today the mall was full of moms who love scrapbooking. kill me., the annotators choose cNo for all the temporal spans. The author of the Tweet was never a participant in scrapbooking despite she witnessed it (presumably shortly before tweeting).

In I love and hate today, I swear I just got a flashback to my sophomore year., the author of the Tweet is a participant in the event swear certainly during tweeting, but not before or after tweeting. The annotators therefore choose cNo for before and after and cYes for during tweet timestamp.

In the tweet Ugh i need a new phone :/ I ay. ! ay is an expression but is recognized as an event. Hence, the annotators gave the label inv for all the temporal spans.

Participation of the author of Tweet Love me when I least deserve it, because that’s when I need it the most in the event need in unknown because he just tweets a random quote but not an actual occurrence of an event. Hence, the annotators gave the label unk for all the temporal spans.

Note that the annotations also provide information regarding event durations, e.g., addicted in example Tweet above is likely ongoing a day after, but talking and swear in have ended and were short events.
4.1. Label Distribution

In figure 4.1 we present overall distribution (column 1) and the label distribution per temporal span (columns 2-6). From the overall distribution $c_{\text{No}}$ is 46.7% and $c_{\text{Yes}}$ is 28.13%, i.e., the annotators are certain that the author of the tweet is or is not a participant in the event. Percentage of $p_{\text{Yes}}$ (11.44%) and $p_{\text{No}}$ (2.93%) are much lower.

People do not usually tweet about events in which they participate while tweeting ($c_{\text{Yes}}$: 33.4% vs $c_{\text{No}}$: 54.5%). People are more likely to tweet about events in which they participate within the last 24 hrs (39.1%) than longer before (26.6%), and after tweeting (24.9% and 16.8%). The per-
centage of $p_{Yes}$ and $p_{No}$ are below 2% for tweet timestamp. Intuitively, it is easier to determine whether somebody participates in the event he tweets about when tweeting rather than before or after.

Percentages for labels that do not indicate event participation ($unk$ and $inv$) are low: $unk$ ranges from 0.91% to 1.91%, and $inv$ is 9.5% for all spans.

4.2. Event Analysis

**Figure 4.2.** Overall distribution of labels per part-of-speech tags of the event.

Figure 4.2 presents the label distribution based on the parts of speech tag of the event for all temporal spans. The percent of $c_{Yes} + p_{Yes}$ is the highest for VBP non-3rd person singular present verb (48.46% vs 28.78% and 31.53%) and hence the author is likely to participate in such events. The author is not likely to participate in events with parts of speech tag VBG gerund or present participle verb (% of $c_{No} + p_{No}$ 61.1%)
In figure 4.3 we portray the label distribution for the top 5 frequent events with their frequency counts in the annotated corpus for all temporal spans. The percent of cYes and pYes is the highest for four of the top five frequent events, and for the event ‘need’ the percent of inv is the highest. e.g., Look, I need less friends more bread, less talk, more head.
CHAPTER 5

EXPERIMENTS AND RESULTS

In this chapter we present in detail the experiments performed and the results obtained on the annotated corpus of event instances described in the previous chapters. We explain the machine learning setup used in our experiments in section 5.1. In section 5.2, we describe the features used for the learning experiments. We discuss the trends observed in the results in section 5.3.

5.1. Experimental Setup

We have 1096 event mentions in the 826 tweets that we annotated. From this we considered 80% (877) of the event mentions for training and remaining 20% (219) for evaluation. We made sure that all the event mentions present in a tweet belong to either train set or test set to avoid bias. We created one instance per event mention and trained five classifiers, one for each temporal span.

We trained Support Vector Machines with RBF kernel using scikit-learn [14]. We tuned the following C and $\gamma$ parameters using 5-fold cross validation on the train set and report precision recall and F1 measures on the test set:

$\gamma$: [1e-1, 1e-2, 1e-3, 1e-4, 1e-5, 1e-6, 1e-7, 1e-8]

C: [1, 10, 100, 1000]

5.2. Features Used

In Table 5.1, we list the features used in our experiments. Features 1 through 3 are event related features. Feature 1 is the event word as it appears in the tweet. Feature 2 is the parts of speech tag of the event word. We obtain the parts of speech tags from T-POS[15]. Feature 3 is the index of the event token in the tweet (token indices start from 0).

Features 4 through 8 are partially inspired by the situational entities project [5]. Feature 4 is the fine tense of the event. Fine tense of the event indicates if the event is in simple past, past continuous, past perfect, past perfect continuous, simple present, present continuous, present perfect, present perfect continuous, future, future continuous, future perfect or future perfect continuous. The fine tense of the event is determined by extracting the event’s verb chain and by set
| Type                      | No. | Description                                                                 |
|---------------------------|-----|-----------------------------------------------------------------------------|
| **Event**                 | 1   | Word form of the event                                                     |
|                           | 2   | Parts of speech tag of the event                                           |
|                           | 3   | Token number of event within the tweet                                     |
| **Situational Entities**  | 4   | Fine tense of the event                                                    |
|                           | 5   | Flag indicating whether event tense is perfect tense                       |
|                           | 6   | Modal type of event, if any                                               |
|                           | 7   | Flag indicating whether event is a reporting verb                          |
|                           | 8   | WordNet lexical file name of event                                         |
| **Context**               | 9   | Position of any pronouns with respect to event: left, right, both or none   |
|                           | 10  | Position of pronouns I, me, we with respect to event: left, right, both or none |
|                           | 11  | Number of outgoing dependencies from event                                 |
|                           | 12  | Flag indicating whether there is a dependency between event and pronouns I, me, or we |
| **Tweet level**           | 13  | No. tokens to the left of the event                                        |
|                           | 14  | No. of tokens to the right of the event                                    |
|                           | 15  | Parts of speech tag of token to the left of the event                      |
|                           | 16  | Parts of speech tag of token to the right of the event                     |

**Table 5.1.** Features to determine whether the author of a tweet is a participant in the events he tweets about.

of rules defined in [10]. Feature 5 is a flag that takes the value true if the fine tense of the event is past perfect, past perfect continuous, present perfect, present perfect continuous, future perfect or future perfect continuous and false otherwise. Feature 6 indicates if the event has a conditional modal or future modal associated with it. Feature 7 takes the value true if the event belongs to a list of reporting verbs that we collected [4] and false otherwise. Feature 8 is the lexical filename of the event lemma in wordnet [12].

Features 9 through 12 capture the context of the event. We indicate the presence of any pronoun to the left or right of the event token in feature 9. If there is a pronoun to the left and right of the event token, feature 9 takes the value both. Feature 10 takes the values left, right or both if there is a pronoun I, me or we to the left, right or left and right of the event token respectively. We obtain the dependency graph of the tweet using Tweeboparser [7]. We represent the number of outgoing dependencies from the event token in feature 11. Feature 12 takes the value true if there is a pronoun I, me or we in the tweet and there is a dependency between the event token and the
pronoun. In tweet level features 13 through 16 we indicate the number of tokens to the left and right of the event token and the parts of speech tags of the token to the left and right of the event token.

![FIGURE 5.1. Dependency graph obtained from tweebo parser for the example tweet](image)

| Tweet:          | #HarryPotter Marathon while I write my papers. |
|-----------------|------------------------------------------------|
| Type            | No. | Name   | Value       |
| Event           | 1   | Event Word | write       |
|                 | 2   | Event POS | VBP         |
| Situational Entities | 3   | Event tense | pres        |
|                 | 4   | Is perfect | FALSE       |
|                 | 5   | Event modal | NULL        |
|                 | 6   | Is reporting verb | FALSE |
|                 | 7   | WN lexical file name | verb.creation |
| Context         | 8   | Position of any pronouns | left |
|                 | 9   | Position of pronouns I, me, we | left |
|                 | 10  | Outgoing dependencies | 2 |
|                 | 11  | Event pronoun dependency flag | FALSE |
| Tweet level     | 12  | # Left tokens | 4 |
|                 | 13  | # Right tokens | 3 |
|                 | 14  | Left POS | PRP |
|                 | 15  | Right POS | PRP$ |

TABLE 5.2. Feature values for an example tweet.

We present an example tweet and its corresponding feature values in Table 5.2. For the event *write* which is the 4th token in the given tweet, the parts of speech tag obtained from T-POS [15] is VBP (non third person singular present verb). The event is in simple present and hence the feature 4 is set to false. The value of feature 5 is NULL since there is no conditional modal or future modal associated with the event. The token *write* is not a reporting verb and belongs to the lexical file *verb.creation* in wordnet. Since there is the pronoun *I* to the left of *write* in the tweet,
the features 8 and 9 take the value left. There are no outgoing dependencies from the token write, so the feature values for 10 is 0 and 11 is FALSE. There are 4 tokens to the left of write and 3 tokens to its right (features 12 and 13). The parts of speech tag of left token I is PRP (personal pronoun) and right token my is PRP$ (possessive pronoun)

5.3. Experimental Results

In this section we present and discuss the results obtained in our learning experiments. We performed learning experiments:

1. Using fine grained labels: cYes, cNo, pYes, pNo, unk and inv.
2. Using coarse grained labels: Yes (combining labels cYes and pYes), No (combining labels cNo and pNo), unk and inv.

5.3.1. Results with fine grained labels

| Baseline | ≥24h Before | <24h Before | tweet timestamp | <24h After | ≥24h After |
|----------|-------------|-------------|----------------|-------------|-------------|
|          | P   | R   | F   | P   | R   | F   | P   | R   | F   | P   | R   | F   | P   | R   | F   | P   | R   | F   |
| cNo      | 0.45 | 1.00 | 0.62 | 0.38 | 1.00 | 0.55 | 0.50 | 1.00 | 0.67 | 0.33 | 1.00 | 0.49 | 0.39 | 1.00 | 0.56 |
| others   | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Avg.     | 0.20 | 0.45 | 0.28 | 0.15 | 0.38 | 0.21 | 0.25 | 0.50 | 0.34 | 0.11 | 0.33 | 0.16 | 0.15 | 0.39 | 0.22 |

**Table 5.3.** Results with most frequent baseline per temporal span using fine grained labels

Most frequent baseline (Table 5.3) predicts the most frequent label cNo for all instances in each temporal span. The results for tweet timestamp temporal span (0.34 F-measure) are higher compared to all other temporal spans.

| Event | ≥24h Before | <24h Before | tweet timestamp | <24h After | ≥24h After |
|-------|-------------|-------------|----------------|-------------|-------------|
|       | P   | R   | F   | P   | R   | F   | P   | R   | F   | P   | R   | F   | P   | R   | F   | P   | R   | F   |
| cYes  | 0.60 | 0.51 | 0.55 | 0.71 | 0.71 | 0.71 | 0.94 | 0.69 | 0.79 | 0.84 | 0.59 | 0.69 | 0.85 | 0.69 | 0.76 |
| pYes  | 0.00 | 0.00 | 0.00 | 0.33 | 0.06 | 0.10 | 0.00 | 0.00 | 0.00 | 0.51 | 0.33 | 0.40 | 0.37 | 0.25 | 0.30 |
| cNo   | 0.56 | 0.85 | 0.67 | 0.54 | 0.74 | 0.63 | 0.69 | 0.93 | 0.79 | 0.45 | 0.79 | 0.58 | 0.48 | 0.74 | 0.59 |
| pNo   | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.14 | 0.11 | 0.12 |
| unk   | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Avg.  | 0.48 | 0.56 | 0.50 | 0.57 | 0.61 | 0.57 | 0.74 | 0.75 | 0.73 | 0.55 | 0.53 | 0.51 | 0.50 | 0.52 | 0.49 |

**Table 5.4.** Results with event features per temporal span using fine grained labels
We trained the classifiers with event features only and the obtained results show good improvement over the baseline (table 5.4) for all the temporal spans. The results with the most frequent label \texttt{cNo} also improved for all the temporal spans. Event features also helped in improving the performance of \texttt{cYes} label. Event features alone gave the best results for $<24$ hrs before and $<24$ hrs after temporal spans.

| Label | $\geq 24$ hrs Before | $<24$ hrs Before | tweet timestamp | $<24$ After | $\geq 24$ After |
|-------|----------------------|------------------|-----------------|------------|---------------|
|       | P  | R  | F  | P  | R  | F  | P  | R  | F  | P  | R  | F  | P  | R  | F  | P  | R  | F  |
| \texttt{cYes} | 0.61 | 0.52 | 0.56 | 0.71 | 0.69 | 0.70 | 0.91 | 0.73 | 0.81 | 0.75 | 0.60 | 0.67 | 0.85 | 0.71 | 0.78 |
| \texttt{pYes} | 0.00 | 0.00 | 0.00 | 0.50 | 0.06 | 0.11 | 0.00 | 0.00 | 0.00 | 0.50 | 0.30 | 0.37 | 0.42 | 0.31 | 0.36 |
| \texttt{cNo} | 0.57 | 0.81 | 0.67 | 0.56 | 0.73 | 0.64 | 0.72 | 0.87 | 0.79 | 0.46 | 0.74 | 0.57 | 0.49 | 0.69 | 0.58 |
| \texttt{pNo} | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.25 | 0.11 | 0.15 |
| \texttt{unk} | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Avg. | 0.47 | 0.55 | 0.50 | 0.58 | 0.60 | 0.57 | 0.74 | 0.74 | 0.73 | 0.52 | 0.51 | 0.49 | 0.51 | 0.53 | 0.51 |

TABLE 5.5. Results with event and situation entities features per temporal span using fine grained labels

When we trained the classifiers using event and situation entities features, we found that the overall performance slightly improved for $\geq 24$ hrs temporal span. The results of the label \texttt{pYes} also improved with respect to the event only for temporal spans $<24$ hrs before, $<24$ hrs after and $\geq 24$ hrs after. The results for the label \texttt{pNo} improved slightly for $\geq 24$ hrs after. This combination of features gave the best performance for the temporal spans $<24$ hrs before and $\geq 24$ hrs after.

| Label | $\geq 24$ hrs Before | $<24$ hrs Before | tweet timestamp | $<24$ After | $\geq 24$ After |
|-------|----------------------|------------------|-----------------|------------|---------------|
|       | P  | R  | F  | P  | R  | F  | P  | R  | F  | P  | R  | F  | P  | R  | F  | P  | R  | F  |
| \texttt{cYes} | 0.71 | 0.55 | 0.62 | 0.69 | 0.69 | 0.69 | 0.91 | 0.73 | 0.81 | 0.73 | 0.56 | 0.60 | 0.85 | 0.69 | 0.76 |
| \texttt{pYes} | 0.00 | 0.00 | 0.00 | 0.50 | 0.06 | 0.11 | 0.00 | 0.00 | 0.00 | 0.55 | 0.40 | 0.46 | 0.35 | 0.31 | 0.33 |
| \texttt{cNo} | 0.57 | 0.84 | 0.68 | 0.55 | 0.69 | 0.61 | 0.73 | 0.89 | 0.80 | 0.42 | 0.61 | 0.49 | 0.50 | 0.65 | 0.56 |
| \texttt{pNo} | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | 0.11 | 0.14 |
| \texttt{unk} | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Avg. | 0.50 | 0.58 | 0.52 | 0.57 | 0.59 | 0.56 | 0.75 | 0.76 | 0.74 | 0.50 | 0.48 | 0.48 | 0.49 | 0.50 | 0.49 |

TABLE 5.6. Results with event, situation entities and context features per temporal span using fine grained labels

When we add context features, best performance is obtained for the temporal spans $\geq 24$ hrs before and tweet timestamp.
The baseline is outperformed by all the feature combinations. Note that while the results with event features are only outperformed for some temporal spans when training with all features, information beyond the event at hand is needed to solve this task. For example, the correct labels for “I love living in NYC” and “I miss living in NYC” are different. Living is an ongoing event in the former examples while past event in the later.

In all the temporal spans and for all the feature combinations, the labels cYes and cNo have higher F-measures compared to other labels. The classifiers are hence able to determine if the author is certainly participating (or not) in the event.

5.3.2. Results with coarse grained labels

| Label | ≥24h Before | <24h Before | tweet timestamp | <24h After | ≥24h After |
|-------|-------------|-------------|-----------------|------------|------------|
|       | P | R | F | P | R | F | P | R | F | P | R | F | P | R | F |
| No    | 0.49 | 1.00 | 0.66 | 0.40 | 1.00 | 0.57 | 0.50 | 1.00 | 0.67 | 0.35 | 1.00 | 0.52 | 0.43 | 1.00 | 0.60 |
| others | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Avg.  | 0.24 | 0.49 | 0.32 | 0.16 | 0.40 | 0.23 | 0.26 | 0.51 | 0.34 | 0.12 | 0.35 | 0.18 | 0.43 | 0.26 |

**TABLE 5.7. Results with most frequent baseline per temporal span using coarse grained labels**

Most frequent baseline (Table 5.7) predicts the most frequent label no for all instances in each temporal span. The results for tweet timestamp temporal span (0.34 F-measure) are higher compared to all other temporal spans.

| Label | ≥24h Before | <24h Before | tweet timestamp | <24h After | ≥24h After |
|-------|-------------|-------------|-----------------|------------|------------|
|       | P | R | F | P | R | F | P | R | F | P | R | F | P | R | F |
| Yes   | 0.70 | 0.47 | 0.57 | 0.70 | 0.68 | 0.69 | 0.91 | 0.67 | 0.77 | 0.79 | 0.64 | 0.71 | 0.68 | 0.57 | 0.62 |
| No    | 0.59 | 0.83 | 0.69 | 0.55 | 0.67 | 0.60 | 0.69 | 0.92 | 0.79 | 0.51 | 0.76 | 0.61 | 0.54 | 0.73 | 0.62 |
| unk   | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Avg.  | 0.61 | 0.62 | 0.59 | 0.60 | 0.62 | 0.61 | 0.75 | 0.75 | 0.73 | 0.64 | 0.63 | 0.62 | 0.57 | 0.59 | 0.57 |

**TABLE 5.8. Results with event features per temporal span using fine grained labels**

We trained the classifiers with event features only and the obtained results show good improvement over the baseline (Table 5.8) for all the temporal spans. The results with the most frequent label no also improved for all the temporal spans. Event features also helped in improving the performance of yes label.
When we trained the classifiers using event and situation entities features, we found that the overall performance slightly improved for tweet timestamp temporal span 5.9. The results of the label yes slightly improved with respect to the event only for temporal span ≥24 hrs before. The results for the label no improved slightly for tweet timestamp. This combination of features gave the best performance for the temporal span tweet timestamp.

**Table 5.9. Results with event and situation entities features per temporal span using coarse grained labels**

| Label | ≥24h Before | <24h Before | tweet timestamp | <24h After | ≥24h After |
|-------|-------------|-------------|-----------------|-----------|-----------|
|       | P  R  F    | P  R  F    | P  R  F        | P  R  F  | P  R  F  |
| Yes   | 0.68 0.53 0.59 | 0.72 0.64 0.68 | 0.95 0.65 0.77 | 0.77 0.57 0.66 | 0.66 0.56 0.62 |
| No    | 0.60 0.73 0.66 | 0.55 0.64 0.59 | 0.72 0.89 0.80 | 0.47 0.70 0.56 | 0.54 0.64 0.58 |
| unk   | 0.00 0.00 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 |
| Avg.  | 0.60 0.60 0.59 | 0.60 0.61 0.60 | 0.77 0.74 0.74 | 0.60 0.58 0.58 | 0.56 0.57 0.56 |

When we add context features the overall performance increases for the temporal spans ≥24 hrs before and <24hrs before 5.10. The performance of the label yes improves for the temporal spans ≥24 hrs before, <24 hrs before and tweet timestamp. The results with the label no improved from all temporal spans except for <24 hrs after. The combination of event, situation entities and context features give the best results with the temporal spans ≥24 hrs before, <before and tweet timestamp for coarse grained labels.

**Table 5.10. Results with event, situation entities and context features per temporal span using coarse grained labels**

| Label | ≥24h Before | <24h Before | tweet timestamp | <24h After | ≥24h After |
|-------|-------------|-------------|-----------------|-----------|-----------|
|       | P  R  F    | P  R  F    | P  R  F        | P  R  F  | P  R  F  |
| Yes   | 0.78 0.51 0.61 | 0.74 0.67 0.70 | 0.95 0.67 0.79 | 0.74 0.58 0.65 | 0.65 0.54 0.59 |
| No    | 0.62 0.80 0.70 | 0.57 0.67 0.62 | 0.73 0.90 0.81 | 0.47 0.66 0.55 | 0.56 0.69 0.62 |
| unk   | 0.00 0.00 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 |
| Avg.  | 0.65 0.63 0.62 | 0.63 0.63 0.63 | 0.75 0.76 0.74 | 0.59 0.58 0.57 | 0.56 0.57 0.56 |

For all the temporal spans, the classifiers are able to predict if the author is a participant in the event (or not) with greater f-measure.
CHAPTER 6

CONCLUSIONS

In this work, we determined if and when the author of a tweet is involved in the event he/she tweets about. We gathered an initial 5017 tweets with 7008 event mentions from various sources. We then processed these event mentions by applying several filters to eliminate events for which determining authors participation is not valid or minimal. After preprocessing we obtained 826 tweets with 1096 event mentions.

For each event mention we asked the question "Is the author of the tweet a participant in the event?" for the following five temporal spans: \( \geq 24 \text{ hrs before tweeting} \), \(<24 \text{ hrs before tweeting} \), during tweet timestamp, \(<24 \text{ hrs after tweeting} \) and \( \geq 24 \text{ hrs after tweeting} \). We specified guidelines to be followed while annotating. We asked two annotators to annotate 10\% of the events and calculated Fleiss kappa values and % disagreements to assess the quality of the annotations. The overall kappa score is 0.70 which is considered as substantial agreement.

We analyzed the annotated corpus with respect to the overall and per temporal span label distributions. From the analysis, we conclude that people do not usually tweet about events in which they participate while tweeting. People are more likely to tweet about events in which they participate within the last 24 hrs than longer before, and after tweeting. It is also easier to determine whether somebody participates in the event he tweets about when tweeting rather than before or after.

We also study the behavior of several events in the corpus and devise that the author is more likely to participate in the events that are on 3rd person singular present verbs.

We elucidate the features used and the standard supervised learning approach to automatically determine if and when the author of the tweet is a participant in the event at hand. We perform experiments with fine-grained labels (cYes, pYes, cNo, pNo, unk and inv) as well as coarse-grained labels(yes, no, unk and inv). From the results obtained we observe:

- For fine-grained labels:
  - Combination of event and situation entities features gave the best performance for
the temporal spans <24 hrs before, <24 hrs after and ≥24 hrs after.

– Combination of an event, situation entities and context features gave the best performance for the temporal spans ≥24 hrs before and tweet timestamp.

– The classifiers are able to determine if the author is certainly participating (or not) in the event.

• For coarse grained labels:

– Event features alone gave the best performance for the temporal spans < and ≥24 hrs after

– Combination of event and situation entities features gave the best performance for the temporal span tweet timestamp.

– The combination of event, situation entities and context features give the best results with the temporal spans ≥24 hrs before, <24 hrs before and tweet timestamp.
[1] Tao Cheng and Thomas Wicks, *Event detection using twitter: A spatio-temporal approach*, PLOS ONE 9 (2014), no. 6, 1–10.

[2] Barbara Di Eugenio, Nick Green, and Rajen Subba, *Detecting life events in feeds from twitter*, Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on, IEEE, 2013, pp. 274–277.

[3] Thomas Dickinson, Miriam Fernandez, Lisa A. Thomas, Paul Mulholland, Pam Briggs, and Harith Alani, *Identifying prominent life events on twitter*, Proceedings of the 8th International Conference on Knowledge Capture (New York, NY, USA), K-CAP 2015, ACM, 2015, pp. 4:1–4:8.

[4] Education first, *Reporting Verbs*, http://www.ef.edu/english-resources/english-grammar/reporting-verbs/.

[5] Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal, *Situation entity types: automatic classification of clause-level aspect*, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Berlin, Germany), Association for Computational Linguistics, August 2016, pp. 1757–1768.

[6] Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith, *A dependency parser for tweets*, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Doha, Qatar), Association for Computational Linguistics, October 2014, pp. 1001–1012.

[7] Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A Smith, *A dependency parser for tweets*, Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, to appear, vol. 4, 2014.

[8] J. Richard Landis and Gary G. Koch, *The measurement of observer agreement for categorical data*, Biometrics 33 (1977), no. 1.

[9] Jiwei Li, Alan Ritter, Claire Cardie, and Eduard Hovy, *Major life event extraction from twitter based on congratulations/condolences speech acts*, Proceedings of the 2014 Conference on
Empirical Methods in Natural Language Processing (EMNLP) (Doha, Qatar), Association for Computational Linguistics, October 2014, pp. 1997–2007.

[10] Sharid Loaiciga, Thomas Meyer, and Andrei Popescu-Belis, English-french verb phrase alignment in europarl for tense translation modeling, The Ninth Language Resources and Evaluation Conference, no. EPFL-CONF-198442, 2014.

[11] André FT Martins, Noah A Smith, and Eric P Xing, Concise integer linear programming formulations for dependency parsing, Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volome 1, Association for Computational Linguistics, 2009, pp. 342–350.

[12] George A. Miller, Wordnet: A lexical database for english, Commun. ACM 38 (1995), no. 11, 39–41.

[13] Olutobi Owoputi, Brendan O’Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith, Improved part-of-speech tagging for online conversational text with word clusters, Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Atlanta, Georgia), Association for Computational Linguistics, June 2013, pp. 380–390.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011), 2825–2830.

[15] Alan Ritter, Sam Clark, Oren Etzioni, et al., Named entity recognition in tweets: an experimental study, Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, pp. 1524–1534.

[16] Alan Ritter, Mausam, Oren Etzioni, and Sam Clark, Open domain event extraction from twitter, Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (New York, NY, USA), KDD ’12, ACM, 2012, pp. 1104–1112.

[17] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo, Earthquake shakes twitter users: Real-
time event detection by social sensors, Proceedings of the 19th International Conference on World Wide Web (New York, NY, USA), WWW ’10, ACM, 2010, pp. 851–860.

[18] Roser Sauri and James Pustejovsky, Factbank: a corpus annotated with event factuality., Language Resources and Evaluation 43 (2009), no. 3, 227–268.

[19] Deyu Zhou, Liangyu Chen, and Yulan He, A simple bayesian modelling approach to event extraction from twitter, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (Baltimore, Maryland), Association for Computational Linguistics, June 2014, pp. 700–705.