

Determining Event Outcomes: The Case of #fail

Srikala Murugan, Dhivya Chinnappa* and Eduardo Blanco

Human Intelligence and Language Technologies Lab

University of North Texas

{srikalamurugan, dhivyainfantchinnappa}@my.unt.edu, eduardo.blanco@unt.edu

Abstract

This paper targets the task of determining event outcomes in social media. We work with tweets containing either #cookingFail or #bakingFail, and show that many of the events described in them resulted in something edible. Tweets that contain images are more likely to result in edible albeit imperfect outcomes. Experimental results show that edibility is easier to predict than outcome quality.

1 Introduction

While the definition of *event* is controversial (Casati and Varzi, 2020; Sprugnoli and Tonelli, 2016), there is general consensus that events occur (or happen, or take place) at a time and location. People share in social media a deluge of information including events they care about. These events range from mundane events such as eating or watching TV to important life events such as getting married and graduating from college (Li et al., 2014). Twitter is one of the most popular social networks with 166 million daily active users (Twitter, 2020).

An important property of events is whether they actually occurred. The literature has studied this property under different terms, e.g., factuality (Saurí and Pustejovsky, 2009; Lee et al., 2015) and veridicality (de Marneffe et al., 2012). Other related tasks have studied the level of commitment a speaker or writer has towards a proposition (Werner et al., 2015; Jiang and de Marneffe, 2019). Assessing the degree to which an event occurred or is believed to be true is critical to make inferences and information extraction. Even when an event is guaranteed to have occurred, however, it is not necessarily the case that the desired outcome came to fruition. For example, people make phone

* Currently at Thomson Reuters. Work done while at University of North Texas.



Figure 1: Tweet discussing a baking event. Despite the presence of the #BakingFail hashtag, the baking was not a complete failure. Indeed, it (most likely) resulted in something edible but visually unappealing.

calls (presumably) to communicate with whoever they are calling. Making the call, however, does not guarantee that the communication took place—the recipient could have not picked up the phone. Some events have fairly clear desired outcomes even if they are not explicitly stated: people make phone calls to communicate, run in elections so that they are elected, etc. The desired outcomes of other events, however, are not so clear: people may plant a tree to help the environment, to provide privacy or shade, or so that it bears fruit.

Like factuality, determining whether an event resulted in the desired outcome is a matter of degree and not a binary decision. In other words, events often do not result in perfect outcomes or complete failures. For example, a phone call may result in communication that is far from perfect because there is background noise or because the call suddenly drops. Consider the tweet in Figure 1. Despite the hashtag #BakingFail, the baking was partially successful: something edible came out of the baking, although it was not visually appealing.

In this paper, we target cooking and baking events that include some form of the hashtag #fail,

and study the degree to which they resulted in their desired outcomes in terms of edibility and quality. The main contributions are: (a) a corpus of 4,000 tweets annotated with event outcome information in two stages: edibility and quality;¹ (b) analysis showing that more information can be extracted from tweets including an image; (c) experimental results showing that determining outcome quality remains a challenge; and (d) error analysis shedding light into the difficulty of the task.

2 Previous Work

The language of social media has been studied from many angles, including applications in the social sciences (Park et al., 2015) and public health (Paul and Dredze, 2011). In the context of emergencies, detecting the first message about a new disaster and information aggregation are important problems (Imran et al., 2015). In this paper, we work with mundane events (cooking and baking) described in one tweet, and study the degree to which they resulted in their desired outcomes.

Event detection from social media has received considerable attention, in particular, pinpointing important life events (Li et al., 2014; Dickinson et al., 2015). Previous research shows that people often tweet about events they do not participate in (Sanagavarapu et al., 2017), targets recurring events (Kunneman and Van den Bosch, 2015), and summarizes tweet streams about TV shows (Andy et al., 2019). The work presented here is not concerned with event detection, our selection criteria virtually guarantees that we only work with tweets about cooking or baking (Section 3).

Determining the degree to which an event results in its desired outcome is distantly related to assessing factuality and other event properties. Previous efforts working with social media target event factuality (Soni et al., 2014), identify controversial events (Popescu and Pennacchiotti, 2010) and credible eyewitnesses (Doggett and Cantarero, 2016), and work with arguably more challenging properties such as rumors (Zubiaga et al., 2015) and credibility (Castillo et al., 2011; Mitra et al., 2017). In the work presented here, we work with factual mundane events whose credibility is undisputed. Lack of factuality or credibility indicates that an event did not occur thus also that the desired outcomes were not achieved. We note, however, that fac-

¹https://github.com/msrikala/Event_outcome.

tual and credible events did not necessarily result in their desired outcomes, as the examples in this paper illustrate with cooking and baking events.

To our knowledge, there are only a few previous works investigating event outcomes from a computational perspective. Outside the social media domain, Velichkov et al. (2019) investigate models to predict the outcome of sports events from interviews conducted shortly before the event. Within social media, Stowe et al. (2018) present models to determine whether people evacuate during a hurricane event from their tweets. Finally, Swamy et al. (2017) present a framework to forecast winners of events (e.g., sports events, elections, awards) by aggregating predictions made by individual users. Our work differs in many respects. First, we work with mundane events (cooking and baking). Second, we investigate a finer-grained characterization of event outcomes beyond binary decisions: edibility and quality. Third, we work with tweets consisting of only text as well as both text and images, and show that the outcomes are easier to determine in the latter—in particular, edibility, both by humans and computational models.

3 Annotating Event Outcomes: Cooking and Baking Events

We create a new corpus of tweets annotated with event outcome information. Initially we set to work with mundane events carried out by regular people and requiring some degree of skill. We explored the following events: driving, gardening, playing sports, singing, playing musical instruments, cooking, and sewing. After manually observing many tweets discussing these events, it became clear that event outcomes are often unknown for events that do not result in concrete outcomes. Additionally, people barely discuss some of the events above unless they result in the expected outcome (e.g., most people talking about driving appear to be good drivers and reach their destinations). We decided to focus on cooking and baking events because (a) they require minimal expertise (i.e., most people can do some cooking); (b) are frequently discussed in social media; and (c) people often discuss the outcome of their cooking and baking in social media, including less than perfect outcomes.

Selecting Tweets. We downloaded 4,000 English tweets describing cooking or baking events using tweepy.² More specifically, we downloaded 2,000

²<https://github.com/tweepy/>

Annotation Task	txt		txt+img	
	%	κ	%	κ
Relevant?	96.3	0.73	98.6	0.66
Outcome edible?	76.7	0.55	79.7	0.60
Outcome quality?	87.5	0.73	87.1	0.76

Table 1: Inter-annotator agreements with tweets consisting of (a) only text and (b) text and an image. We present the raw agreements (%) and Cohen’s κ .

tweets containing the *#cookingfail* hashtag and 2,000 tweets containing the *#bakingfail* hashtag. Half of the tweets in each category consisted of only text, and the other half consisted of text and an image. As we shall see, it is common to find tweets that talk about cooking and baking failures despite an edible outcome resulted from them, especially when the tweet includes an image.

Annotation Guidelines. Dictionaries define cooking as “prepare food for eating [...]”, and baking as “cook by dry heat especially in an oven” (Merriam-Webster, 2003). Thus the desired outcome of cooking or baking events is to create something edible. Our event outcome annotation guidelines for cooking and baking events go beyond this binary distinction and include three steps.

The first step is to identify **relevant** tweets, which we define as tweets that describe a cooking or baking event involving the author. Annotators choose from the following labels for relevancy:

- *yes*: the tweet is relevant; or
- *no*: the tweet is not relevant.

The majority of the selected tweets are relevant; exceptions include references to cooking shows.

The second step is to identify whether the cooking or baking event resulted in something **edible**. Annotators choose from the following labels:

- *yes*: the cooking or baking event resulted in something edible; or
- *no*: the cooking or baking event did not result in something edible.

We define edible outcomes as outcomes of cooking or baking events that a reasonable person would eat rather than toss in the trash. Edible outcomes need not be perfect or even what a cook intended to make, they only need to be *edible*.

The third step is to identify the **quality** of edible outcomes. After pilot annotations, we decided to let annotators choose among the following labels:

- *as expected*: the cooking or baking event resulted in the expected food or dish, and there

Label	txt	txt+img
Relevant?		
% <i>yes</i>	94.9	90.0
% <i>no</i>	5.1	10.0
Outcome edible?		
% <i>yes</i>	31.8	59.7
% <i>no</i>	68.2	40.3
Outcome quality?		
% <i>as expected</i>	6.4	6.2
% <i>partial success</i>	57.8	72.6
% <i>alternative</i>	16.3	4.3
% <i>unknown</i>	19.5	16.9

Table 2: Label distribution in the tweets consisting of (a) only text and (b) text and an image.

is nothing wrong with it.

- *partial success*: the cooking or baking event resulted in the expected food or dish, but something went wrong: it may be visually unappealing or partially burnt, it may have resulted in less portions than expected, etc.
- *alternative*: the cooking or baking event resulted in an alternative food or dish than the one the cook originally intended.
- *unknown*: I cannot choose any of the other three labels, there is not enough information.

While perfection is hard to achieve, one could consider outcomes annotated *as expected* to be perfect. Outcomes annotated *partial success* or *alternative*, on the other hand, are imperfect. The former results in the expected outcome with some flaw, and the latter in another outcome altogether (e.g., baking cookies and ending up with biscuits).

All annotations were made with respect to the cooking or baking event up to the point the tweet was published. For example, the outcome of a tweet describing a *baking cake* event and mentioning that the oven tripped a circuit breaker would be annotated *not edible* despite it is possible that the baking was successful after resetting the breaker.

Annotation Process and Agreements. Annotations were done in-house by two graduate students. Both of them annotated 15% of tweets in each group (*#cookingfail* or *#bakingfail*, only text or text and image). Table 1 shows the inter-annotator agreements. Cohen’s κ coefficients (Cohen, 1960) range between 0.55 and 0.76, which is considered *substantial*—above 0.80 is considered nearly *perfect* (Artstein and Poesio, 2008).

We note that (a) κ coefficients for both edibility and quality are slightly higher when tweets consist

Tweet with only text	Annotations		
	relevant?	edible?	quality?
1: Eating crumpets and watching master chef #cookingfail	<i>no</i>	<i>n/a</i>	<i>n/a</i>
2: Oh Bugger. My oklava won't fit in my suitcase... #bakingfail	<i>no</i>	<i>n/a</i>	<i>n/a</i>
3: Right, if I wanna cook the appliances need to be plugged in #cookingfail	<i>yes</i>	<i>no</i>	<i>n/a</i>
4: It's been so long since I've made cupcakes I forgot how to load my frosting gun. :x #BakingFail It's all good now. Cupcakes are frosted...	<i>yes</i>	<i>yes</i>	<i>as expected</i>
5: So the plan was to make Oreo Brownies... I wouldn't quite call it that but still taste pretty good #bakingfail	<i>yes</i>	<i>yes</i>	<i>partial success</i>
6: I tried to make an omelet. It turned into scrambled eggs. #cookingfail	<i>yes</i>	<i>yes</i>	<i>alternative</i>
7: Today's dinner so did not go as planned but I guess the important thing is the kids are fed. #cookingfail	<i>yes</i>	<i>yes</i>	<i>unknown</i>

Table 3: Annotation examples of tweets with only text. Relevancy indicates whether the tweet is about cooking or baking. Edibility and quality only applies to tweets describing relevant and edible events respectively.

1:	2:	3:
relevant? edible? quality?	relevant? edible? quality?	relevant? edible? quality?
<i>no</i> <i>n/a</i> <i>n/a</i>	<i>yes</i> <i>no</i> <i>n/a</i>	<i>yes</i> <i>yes</i> <i>as expected</i>
4:	5:	6:
relevant? edible? quality?	relevant? edible? quality?	relevant? edible? quality?
<i>yes</i> <i>yes</i> <i>prtl. success</i>	<i>yes</i> <i>yes</i> <i>alternative</i>	<i>yes</i> <i>yes</i> <i>unknown</i>

Table 4: Annotation examples of tweets with both text and images. Relevancy indicates whether the tweet is about cooking or baking. Edibility and quality only applies to tweets describing relevant and edible events respectively.

of both text and images, and (b) our agreements are on par or better than previous work working with social media data (Holgate et al., 2018).

4 Corpus Analysis

Table 2 provides the label frequency for each annotation task. The majority of the 4,000 tweets selected are about cooking or baking (94.9% and 90.0%). Despite they contain the hashtag #cookingfail or #bakingfail, a substantial amount of tweets consisting of only text resulted in an edible outcome (31.8%), and this is true for the majority (59.7%) of tweets consisting of text and an image.

Regarding quality, most cooking and baking events resulted in the expected dish with some flaw (*partial success*: 57.8% and 72.6%). Additionally, people are more likely to share a picture if the cooking or baking event was a *partial success* rather than resulted in an *alternative* outcome.

4.1 Examples

We present examples of all labels using tweets consisting of only text in Table 3. Example (1) does not discuss cooking by the author of the tweet (relevant: *no*), and in Example (2) it is unclear: oklava is a kitchen utensil but it appears the author is getting

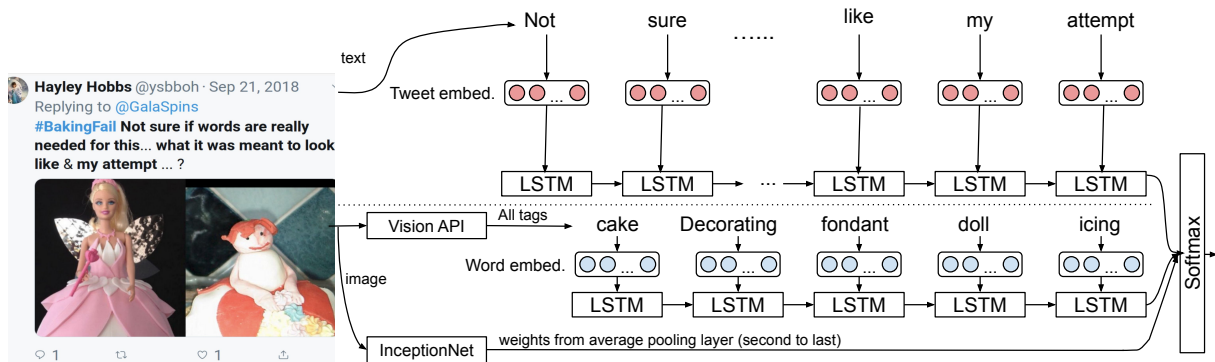


Figure 2: Neural network architecture to predict whether a cooking or baking event resulted in an edible outcome, and if so, the event quality (*as expected*, *partial success*, *alternative* or *unknown*). We include a text component (above dotted line) and two image components (below dotted line).

ready to travel. Unplugged appliances will result in an inedible outcome (Example (3)), and sometimes baking failures refer to some setback that only delays the expected outcome (Example (4)). Examples 5–7 are more nuanced. In Example (5), the outcome had some flaw but was edible (*partial success*), and in Example (6) the author ended up with scrambled eggs while trying to make an omelet. Finally, the outcome in Example (7) is *unknown* because it is unclear how the kids were fed—it is possible that the family ended up ordering takeout food.

Table 4 presents examples with tweets consisting of text and an image. The rationale for the annotations is similar. We note that both the text and image are necessary to annotate correctly. Indeed, the bottom left cupcake in the picture in Example (3) of Table 4 could be misinterpreted as a less than perfect outcome, but the text clearly indicates that they were as good as it gets. Similarly, the text are critical in Examples (4) and (5).

5 Experiments and Results

We experiment with models to predict outcome edibility (*yes* or *no*) and outcome quality (*as expected*, *partial success*, *alternative* or *unknown*). We split the tweets into train (80%) and test (20%) splits, and report results evaluating in the test split with (a) the tweets consisting of only text and (b) tweets consisting of both text and an image.

Baselines. We work with the majority baseline (edibility: always *no* (only text) or *yes* (text and image), quality: always *partial success* for all tweets) and a supervised baseline using Logistic Regression. The Logistic Regression model uses bag-of-words features and only considers the text in tweets as input—it disregards the image if tweets contain

one. We use the implementation in the scikit-learn machine learning Python package (Pedregosa et al., 2011) with default parameters, which in turn uses the LIBLINEAR library (Fan et al., 2008).

Neural Network Architecture. The neural network is inspired by our previous work (Chinnappa et al., 2019) and Cai et al. (2019). It includes two components: one for the text and another one for the image (above and below dotted line in Figure 2). The first component is a basic LSTM (Hochreiter and Schmidhuber, 1997) with 200 units which takes as input the text in the tweet. We lower case tokens and transform them into their GloVe embeddings (Pennington et al., 2014) pretrained with Twitter data (300 dimensions).³

The image component consists of two parts. The first part is another LSTM with 200 units that takes as input the tags automatically extracted from the image by the Google Cloud Vision API.⁴ Note that the tags are an additional text input, and that tags may be more than one word (e.g., chocolate cake), so the LSTM allows us to encode the sequence of tags (which has variable length). Additionally, the word embeddings (GloVe embeddings pre-trained with CommonCrawl) allow us to leverage a distributional representation of tags, including those not seen during training. The second part uses the pre-trained InceptionNet network (Szegedy et al., 2015) in order to extract a representation of the image. More specifically, we use the weights from the average pool layer (second to last).

We implement the neural network with the Keras API (Chollet et al., 2015) and TensorFlow backend (Abadi et al., 2015).

³Available at <https://nlp.stanford.edu/projects/glove/>.

⁴<https://cloud.google.com/vision>

Task and Labels	Maj. Baseline			Log. Regression			NN, only text		
	P	R	F1	P	R	F1	P	R	F1
Outcome is edible?									
<i>yes</i>	0.00	0.00	0.00	0.51	0.44	0.47	0.58	0.50	0.54
<i>no</i>	0.61	1.00	0.76	0.67	0.73	0.70	0.78	0.83	0.81
Weighted Avg.	0.37	0.61	0.46	0.61	0.62	0.61	0.72	0.73	0.72
Outcome quality?									
<i>as expected</i>	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.14	0.20
<i>partial success</i>	0.58	1.00	0.73	0.64	0.80	0.71	0.64	0.80	0.71
<i>alternative</i>	0.00	0.00	0.00	0.44	0.41	0.42	0.46	0.35	0.40
<i>unknown</i>	0.00	0.00	0.00	0.33	0.19	0.24	0.31	0.19	0.24
Weighted Avg.	0.33	0.58	0.42	0.51	0.57	0.53	0.52	0.57	0.53

Table 5: Results obtained with the tweets consisting of only text.

Task and Labels	Maj. Baseline			Log. Regression			NN, text + imgs		
	P	R	F1	P	R	F1	P	R	F1
Outcome is edible?									
<i>yes</i>	0.60	1.00	0.75	0.67	0.76	0.72	0.74	0.81	0.77
<i>no</i>	0.00	0.00	0.00	0.55	0.44	0.49	0.66	0.56	0.60
Weighted Avg.	0.36	0.60	0.46	0.62	0.63	0.62	0.70	0.71	0.70
Outcome quality?									
<i>as expected</i>	0.00	0.00	0.00	0.40	0.17	0.24	0.25	0.08	0.12
<i>partial success</i>	0.63	1.00	0.77	0.64	0.82	0.72	0.66	0.86	0.75
<i>alternative</i>	0.00	0.00	0.00	0.12	0.05	0.07	0.25	0.14	0.18
<i>unknown</i>	0.00	0.00	0.00	0.32	0.21	0.26	0.35	0.17	0.23
Weighted Avg.	0.40	0.63	0.49	0.51	0.57	0.53	0.53	0.60	0.54

Table 6: Results obtained with the tweets consisting of text and images.

5.1 Experimental Results

Tweets with only Text. Table 5 shows the results with tweets consisting of only text. Logistic regression outperforms the majority baseline (weighted F1: 0.61 vs. 0.46). The neural network (only the text component), despite its simplicity, outperforms logistic regression with all labels (F1: 0.72).

Predicting outcome quality is harder. Logistic regression and the neural network obtain the same weighted F1 (0.53) and outperform the majority baseline (F1: 0.42). All the models obtain F1s below 0.50 for all labels except *partial success*, which is the most frequent label.

Tweets with Text and Images. Table 6 shows the results with tweets consisting of text and images. Regarding outcome edibility, we observe a similar pattern as before, but this time the *yes* label (the most frequent) obtains a higher F1 (0.77 vs. 0.60). The neural network (text and image components) outperforms logistic regression predicting outcome edibility (F1: 0.70 vs. 0.62), but not predicting

outcome quality (F1: 0.54 vs. 0.53).

Additional Experiments and Results Additional experimental results are presented in the appendix section. First, we also carry out several ablation experiments in order to check whether the different components of the network are needed. The results show that the full network is beneficial working with tweets consisting of text and images. Overall F1s for outcome edibility with selected components are 0.62 (only text component) 0.63 (only InceptionNet weights), 0.67 (only tags from Vision API and the LSTM to encode them), and 0.66 (full image component). The full network (text and image components) obtains 0.70 F1 (Table 6). For outcome quality, the differences are smaller and all components obtain similar results (F1: 0.50–0.52 vs. 0.53) except using only InceptionNet (0.47).

We also experiment with an alternative set of classes for outcome quality. Specifically, we merge *partial success* and *alternative* as these two labels

Error Type	%	Tweet with only text	Gold	Pred.
Outcome edible?				
World knowledge	54	Went to make banana bread only to discover I have 1 raw egg, 1 hard boiled. CRAP! #bakingfail	<i>no</i>	<i>yes</i>
Human error	15	That moment when you set water on the stove to boil and turn on the wrong burner and walk away. #cookingfail	<i>no</i>	<i>yes</i>
Intricate text	15	“What’s that smell? It smells like eggs... Now it smells like burning...” Oh, wait - it’s me! I was making something, wasn’t I? #cookingfail	<i>no</i>	<i>yes</i>
Alternative outcome	7	Tried to make an omlette turned into scrambled egg.on toast #tastey #cookingfail	<i>yes</i>	<i>no</i>
Outcome quality?				
Word knowledge	41	Left the 15 year old in charge of cooking jacket potatoes and beans for lunch. He put the beans on at the same time as potatoes #cookingfail	<i>partial success</i>	<i>unknown</i>
Lacks information	35	Just made toffee apples with the kids for tea. Now have 4 bowls, 3 spoons & 1 table covered in welded on toffee #bakingfail #puddingsuccess	<i>unknown</i>	<i>partial success</i>
Other	24	So, this chicken with real chicken seasoning sure tastes better than the garlic powder I accidentally used last week. Lol! #cookingfail	<i>as expected</i>	<i>partial success</i>

Table 7: Most frequent error types with tweets consisting of only text. Pred. indicates the predicted label from the best performing model (NN only text, Table 5).

indicate unexpected (but edible) outcomes. The results are as one would expect: it is easier to predict three instead of four labels. The baseline, however, also obtains better results, and in fact both logistic regression and the neural network yield lower relative improvements with respect to the baseline.

6 Error Analysis

We identify the most common error types made by the best model (*NN, only text* and *NN, text + imgs*) after manually analyzing 100 errors.

Tweets with Only Text. Table 7 presents the most frequent error types with tweets consisting of only text. Regarding outcome edibility, the most common type (54%) is the need for *world knowledge*—primarily related to cooking. In the example, annotators had no issue realizing that hard boiled eggs cannot be used for baking, but the model, unsurprisingly, failed to do so. The next two most common errors are *human errors* and *intricate text* (15% each). The former refers to instances in which a human makes the wrong measurement, fails to properly operate appliances, or is otherwise responsible for an inedible outcome. The latter are tweets in which complex reasoning in addition to knowledge about cooking is required. Finally, 7% of errors occurred predicting inedible outcome when in reality

an *alternative (and edible) outcome* resulted from the cooking or baking.

Regarding outcome quality, we identify two major error types. The most common (41%) is also *world knowledge*. In the example, one must know that potatoes and beans have different cooking times; note that the text does not give any explicit cue about the quality of the resulting dish. A substantial amount of errors occur with tweets whose text lacks information to establish the outcome quality (gold: *unknown*). In this case, the model tends to predict the majority label, *partial success*. Finally, the remaining errors (26%) are due to other reasons. In the example, the *#cooking-fail* refers to a past cooking (last week), not the one that occurred shortly before tweeting.

Tweets with Text and Images. Table 8 presents the most common error types with tweets consisting of text and an image. Compared to tweets consisting of only text, we observe that the picture is often critical to make the right prediction—even if the text is long. World knowledge is not a common error type, suggesting that people use pictures for rather explicit outcomes—assuming one can properly interpret the picture. Although we did not anticipate this insight prior to the error analysis, it is to a large extent unsurprising: it is rather hard to

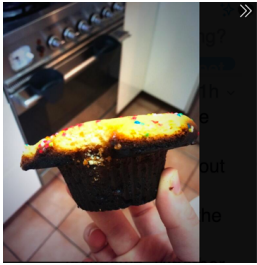
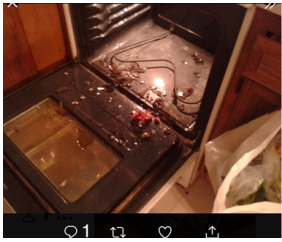


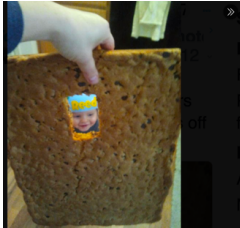

Image is key (25%)	Human error (20%)	Alternative outcome (20%)
 <p>Siob... @Dor... Would you like some cupcake with your charcoal? #bakingfail 4:52 PM · Apr 12,</p>	 <p>Bet... @ca... Oven 1, me 0. Never add liquid to a hot glass dish. #cookingfail</p>	 <p>Meggles @babygl... Christmas cupcakes turned out more like biscuits, I'm sure @gazzzwood will still eat them though! #bakingfail</p>
Gold edible: <i>no</i> Pred. edible: <i>yes</i>	Gold edible: <i>no</i> Pred. edible: <i>yes</i>	Gold edible: <i>yes</i> Pred. edible: <i>no</i>
Lacks information (28%)	Image is key (26%)	Human error (7%)
 <p>@all... #Doh...#bakingfail 8:23 AM · Nov 2, 2017 · Twitter for iPhone</p>	 <p>Anna Hart @annar... Ahh...so this is what happens to cookie bars when a little boy shuts off the timer#bakingfail 3:17 PM · Dec 25</p>	 <p>@jo... Sour cream donuts but forgot to add the leavening. #cookingFail</p>
Gold quality: <i>unknown</i> Pred. quality: <i>partial success</i>	Gold quality: <i>partial success</i> Pred. quality: <i>unknown</i>	Gold quality: <i>partial success</i> Pred. quality: <i>unknown</i>

Table 8: Most frequent error types with tweets consisting of text and images (top: outcome edibility, bottom: outcome quality). Pred. indicates the predicted label from the best performing model (NN text+img, Table 6).

depict world knowledge in a picture.

Regarding outcome edibility (top three examples in Table 8), a common source of errors (25%) is with tweets in which the *image is key*. For example, the text in the first example alone does not make it clear what *charcoal* refers to, but the picture clearly shows that the cupcake is partially burnt. The second cause of errors (20%) is due to *human errors* (mismeasurements, improper use of appliances, etc.) In the second example, the picture is also important but the text alone gives a clue that the cook lost the battle) against the oven (*Oven 1, me 0*), thus we consider it a human error. The third error type (20%) is also shared with the tweets consisting of only text: the model struggles identifying edible outcomes that were not anticipated (i.e., alternative (and edible) outcomes).

Regarding outcome quality, we observe two error types covering over half of the errors and a long tail of additional types. First, some tweets *lack information* in the text and image (28% of errors) to determine the outcome quality (gold: *unknown*), and the model tends to predict the majority label

(*partial success*). Second, the *image is key* in 26% of errors, as illustrated with in the second example. In this example, the event outcome (edibility and quality) is very ambiguous without looking at the picture. Finally, we also identified that the model struggles to identify *partial success* when cooks make some mistake (*human error*, 7% of all errors). In the third example, the cook forgot an ingredient but doing so did not result in a complete failure.

7 Conclusions

Factual and credible events do not necessarily result in their expected outcomes. In this paper, we target outcomes of cooking and baking events from social media. Specifically, we determine whether something edible resulted from them, and also the outcome quality (*as expected*, *partial success* or *alternative*). An annotation effort with 4,000 tweets consisting of either only text or text and an image shows that people often use the hashtag *#cookingFail* or *#bakingFail* when the cooking did not result in a complete failure. Indeed, the outcome is often edible albeit not perfect, especially if the tweet includes an image (59.7 vs. 31.8%).

We believe that a similar approach could be used to assess outcomes of other events. For example, taking exams and going to the grocery store usually have clear expected outcomes: to pass the exam and to buy something. Taking an exam or going shopping (factuality is not in question here), however, does not guarantee that the expected outcomes become a reality (e.g., people take exams and fail them). One may be able to determine not only whether instances of these events occurred, but also if they resulted in the desired outcomes.

Acknowledgements

This material is based in part upon work supported by the National Science Foundation under Grant No. 1820666. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

- Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. *TensorFlow: Large-scale machine learning on heterogeneous systems*. Software available from tensorflow.org.
- Anietie Andy, Derry Tanti Wijaya, and Chris Callison-Burch. 2019. *Winter is here: Summarizing twitter streams related to pre-scheduled events*. In *Proceedings of the Second Workshop on Storytelling*, pages 112–116, Florence, Italy. Association for Computational Linguistics.
- Ron Artstein and Massimo Poesio. 2008. *Survey article: Inter-coder agreement for computational linguistics*. *Computational Linguistics*, 34(4):555–596.
- Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. *Multi-modal sarcasm detection in twitter with hierarchical fusion model*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.
- Roberto Casati and Achille Varzi. 2020. *Events*. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, summer 2020 edition. Metaphysics Research Lab, Stanford University.
- Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. *Information credibility on twitter*. In *Proceedings of the 20th International Conference on World Wide Web, WWW '11*, page 675–684, New York, NY, USA. Association for Computing Machinery.
- Dhivya Chinnappa, Srikala Murugan, and Eduardo Blanco. 2019. *Extracting possessions from social media: Images complement language*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 663–672, Hong Kong, China. Association for Computational Linguistics.
- François Chollet et al. 2015. *Keras*. <https://github.com/fchollet/keras>.
- Jacob Cohen. 1960. *A coefficient of agreement for nominal scales*. *Educational and Psychological Measurement*, 20(1):37–46.
- Thomas Dickinson, Miriam Fernandez, Lisa A. Thomas, Paul Mulholland, Pam Briggs, and Harith Alani. 2015. *Identifying prominent life events on twitter*. In *Proceedings of the 8th International Conference on Knowledge Capture, K-CAP 2015*, New York, NY, USA. Association for Computing Machinery.
- Erika Doggett and Alejandro Cantarero. 2016. *Identifying eyewitness news-worthy events on twitter*. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 7–13, Austin, TX, USA. Association for Computational Linguistics.
- Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. *LIBLINEAR: A library for large linear classification*. *Journal of Machine Learning Research*, 9:1871–1874.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. *Long short-term memory*. *Neural Comput.*, 9(8):1735–1780.
- Eric Holgate, Isabel Cachola, Daniel Preoțiuc-Pietro, and Junyi Jessy Li. 2018. *Why swear? analyzing and inferring the intentions of vulgar expressions*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4405–4414, Brussels, Belgium. Association for Computational Linguistics.
- Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. *Processing social media messages in mass emergency: A survey*. *ACM Comput. Surv.*, 47(4).
- Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. *Evaluating BERT for natural language inference: A case study on the CommitmentBank*. In *Proceedings*

- of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6086–6091, Hong Kong, China. Association for Computational Linguistics.
- Florian Kunneman and Antal Van den Bosch. 2015. **Automatically identifying periodic social events from twitter**. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 320–328, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.
- Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. **Event detection and factuality assessment with non-expert supervision**. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1648, Lisbon, Portugal. Association for Computational Linguistics.
- Jiwei Li, Alan Ritter, Claire Cardie, and Eduard Hovy. 2014. **Major life event extraction from twitter based on congratulations/condolences speech acts**. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1997–2007, Doha, Qatar. Association for Computational Linguistics.
- Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. **Did it happen? the pragmatic complexity of veridicality assessment**. *Computational Linguistics*, 38(2):301–333.
- Merriam-Webster. 2003. *Merriam-Webster's Collegiate Dictionary*, 11th edition. Merriam-Webster.
- Tanushree Mitra, Graham P. Wright, and Eric Gilbert. 2017. **A parsimonious language model of social media credibility across disparate events**. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '17*, page 126–145, New York, NY, USA. Association for Computing Machinery.
- Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. **Automatic personality assessment through social media language**. *Journal of personality and social psychology*, 108(6):934.
- Michael J Paul and Mark Dredze. 2011. **You are what you tweet: Analyzing twitter for public health**. In *Fifth International AAAI Conference on Weblogs and Social Media*.
- 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. 2011. **Scikit-learn: Machine learning in Python**. *Journal of Machine Learning Research*, 12:2825–2830.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. **Glove: Global vectors for word representation**. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Ana-Maria Popescu and Marco Pennacchiotti. 2010. **Detecting controversial events from twitter**. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10*, page 1873–1876, New York, NY, USA. Association for Computing Machinery.
- Krishna Chaitanya Sanagavarapu, Alakananda Vempala, and Eduardo Blanco. 2017. **Determining whether and when people participate in the events they tweet about**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 641–646, Vancouver, Canada. Association for Computational Linguistics.
- Roser Saurí and James Pustejovsky. 2009. **Factbank: a corpus annotated with event factuality**. *Language resources and evaluation*, 43(3):227.
- Sandeep Soni, Tanushree Mitra, Eric Gilbert, and Jacob Eisenstein. 2014. **Modeling factuality judgments in social media text**. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 415–420, Baltimore, Maryland. Association for Computational Linguistics.
- Rachele Sprugnoli and Sara Tonelli. 2016. **One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective**. *Natural Language Engineering*, 23:1–22.
- Kevin Stowe, Jennings Anderson, Martha Palmer, Leysia Palen, and Ken Anderson. 2018. **Improving classification of twitter behavior during hurricane events**. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 67–75, Melbourne, Australia. Association for Computational Linguistics.
- Sandesh Swamy, Alan Ritter, and Marie-Catherine de Marneffe. 2017. **“i have a feeling trump will win.....”:** Forecasting winners and losers from user predictions on twitter. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1583–1592, Copenhagen, Denmark. Association for Computational Linguistics.
- Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. **Going deeper with convolutions**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.
- Twitter. 2020. **Q1 2020 Letter to Shareholders**. <https://investor.twitterinc.com>.

com/financial-information/quarterly-results/default.aspx.
Accessed: May 22, 2020.

obtains better results, but note that the majority baseline also obtains better results.

Boris Velichkov, Ivan Koychev, and Svetla Boytcheva. 2019. [Deep learning contextual models for prediction of sport event outcome from sportsman’s interviews](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1240–1246, Varna, Bulgaria. INCOMA Ltd.

Gregory Werner, Vinodkumar Prabhakaran, Mona Diab, and Owen Rambow. 2015. [Committed belief tagging on the factbank and LU corpora: A comparative study](#). In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, pages 32–40, Denver, Colorado. Association for Computational Linguistics.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. 2015. [Crowdsourcing the annotation of rumourous conversations in social media](#). In *Proceedings of the 24th International Conference on World Wide Web, WWW ’15 Companion*, page 347–353, New York, NY, USA. Association for Computing Machinery.

A Additional Experimental Results

Tables 9–12 present additional experimental results:

- Table 9 shows the results with tweets consisting of both text and images with several ablations of the full neural network. These ablation experiments complement the results in Table 6 of the paper and detail the results with the following parts of the network:
 - Only the text component (NN, only text),
 - Only the InceptionNet weights (NN, img. IN),
 - Only the LSTM that takes as input the tags identified with the Vision API (NN, img. tags), and
 - Only the full image component (NN, IN + tags).

These results show that the the full network described in the main paper obtains better results than any of the individual components: text component only or image component only (either part of the image or both).

- Tables 10, 11, 12 complement Tables 9, 5 and 6 respectively. They compare the results obtained predicting outcome quality using 4 and 3 labels (merging *partial success* and *alternative*). Unsurprisingly, predicting three labels

Task and Labels	NN, only text			NN, img. IN			NN, img. tags			NN, IN + tags		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Outcome is edible?												
<i>yes</i>	0.68	0.72	0.70	0.69	0.71	0.70	0.71	0.77	0.74	0.70	0.77	0.74
<i>no</i>	0.53	0.49	0.51	0.54	0.50	0.52	0.60	0.53	0.56	0.59	0.50	0.54
Weighted Avg.	0.62	0.63	0.62	0.63	0.63	0.63	0.67	0.67	0.67	0.66	0.67	0.66
Outcome quality?												
<i>as expected</i>	0.20	0.08	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
<i>partial success</i>	0.64	0.78	0.70	0.61	0.85	0.71	0.63	0.76	0.69	0.64	0.78	0.70
<i>alternative</i>	0.20	0.14	0.16	0.14	0.05	0.07	0.29	0.23	0.26	0.20	0.09	0.13
<i>unknown</i>	0.31	0.19	0.24	0.11	0.05	0.07	0.25	0.17	0.20	0.29	0.24	0.26
Weighted Avg.	0.50	0.55	0.52	0.42	0.55	0.47	0.48	0.54	0.50	0.48	0.55	0.51

Table 9: Results obtained with tweets consisting of text and images using several components of the proposed neural network. IN refers to features extracted from the pretrained InceptionNet network, and tags refers to the LSTM taking as input the tags from the Google Vision API.

Task and Labels	NN, only text			NN, img. IN			NN, img. tags			NN, IN + tags		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Outcome quality?												
<i>as expected</i>	0.20	0.08	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
<i>partial success</i>	0.64	0.78	0.70	0.61	0.85	0.71	0.63	0.76	0.69	0.64	0.78	0.70
<i>alternative</i>	0.20	0.14	0.16	0.14	0.05	0.07	0.29	0.23	0.26	0.20	0.09	0.13
<i>unknown</i>	0.31	0.19	0.24	0.11	0.05	0.07	0.25	0.17	0.20	0.29	0.24	0.26
Weighted Avg.	0.50	0.55	0.52	0.42	0.55	0.47	0.48	0.54	0.50	0.48	0.55	0.51
Outcome quality?												
<i>as expected</i>	0.33	0.08	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
<i>partial success or alternative</i>	0.77	0.89	0.83	0.73	0.82	0.77	0.75	0.88	0.81	0.76	0.86	0.80
<i>unknown</i>	0.37	0.24	0.29	0.25	0.21	0.23	0.25	0.12	0.16	0.32	0.24	0.27
Weighted Avg.	0.67	0.71	0.68	0.59	0.65	0.62	0.60	0.67	0.63	0.63	0.68	0.65

Table 10: Results obtained training and testing with four and three labels for outcome quality. These results are obtained with tweets consisting of text and images using several components of the proposed neural network. IN refers to features extracted from the pretrained InceptionNet network, and tags refers to the LSTM taking as input the tags from the Google Vision API.

Task and Labels	Maj. Baseline			Log. Regression			NN, only text		
	P	R	F1	P	R	F1	P	R	F1
Outcome quality?									
<i>as expected</i>	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.14	0.20
<i>partial success</i>	0.58	1.00	0.73	0.64	0.80	0.71	0.64	0.80	0.71
<i>alternative</i>	0.00	0.00	0.00	0.44	0.41	0.42	0.46	0.35	0.40
<i>unknown</i>	0.00	0.00	0.00	0.33	0.19	0.24	0.31	0.19	0.24
Weighted Avg.	0.33	0.58	0.42	0.51	0.57	0.53	0.52	0.57	0.53
Outcome quality?									
<i>as expected</i>	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.14	0.20
<i>partial success or alternative</i>	0.74	1.00	0.85	0.76	0.94	0.84	0.79	0.86	0.82
<i>unknown</i>	0.00	0.00	0.00	0.44	0.19	0.27	0.39	0.33	0.36
Weighted Avg.	0.54	0.74	0.62	0.65	0.73	0.67	0.68	0.71	0.69

Table 11: Results obtained training and testing with four or three labels for outcome quality. These results are obtained with tweets consisting of only text. IN refers to features extracted from the pretrained InceptionNet network, and tags refers to the LSTM taking as input the tags from the Google Vision API.

Task and Labels	Maj. Baseline			Log. Regression			NN, text + imgs		
	P	R	F1	P	R	F1	P	R	F1
Outcome quality?									
<i>as expected</i>	0.00	0.00	0.00	0.40	0.17	0.24	0.25	0.08	0.12
<i>partial success</i>	0.63	1.00	0.77	0.64	0.82	0.72	0.66	0.86	0.75
<i>alternative</i>	0.00	0.00	0.00	0.12	0.05	0.07	0.25	0.14	0.18
<i>unknown</i>	0.00	0.00	0.00	0.32	0.21	0.26	0.35	0.17	0.23
Weighted Avg.	0.40	0.63	0.49	0.51	0.57	0.53	0.53	0.60	0.54
Outcome quality?									
<i>as expected</i>	0.00	0.00	0.00	1.00	0.08	0.15	0.17	0.08	0.11
<i>partial success or alternative</i>	0.74	1.00	0.85	0.76	0.91	0.83	0.76	0.93	0.84
<i>unknown</i>	0.00	0.00	0.00	0.35	0.19	0.25	0.31	0.10	0.15
Weighted Avg.	0.54	0.74	0.63	0.69	0.71	0.67	0.66	0.71	0.67

Table 12: Results obtained training and testing with four or three labels for outcome quality. These results are obtained with tweets consisting of text and images. IN refers to features extracted from the pretrained InceptionNet network, and tags refers to the LSTM taking as input the tags from the Google Vision API.