“i have a feeling trump will win.................”: Forecasting Winners and Losers from User Predictions on Twitter

Sandesh Swamy, Alan Ritter
Computer Science & Engineering
The Ohio State University
Columbus, OH
swamy.14@osu.edu, aritter@cse.ohio-state.edu

Marie-Catherine de Marneffe
Department of Linguistics
The Ohio State University
Columbus, OH
mcdm@ling.ohio-state.edu

Abstract

Social media users often make explicit predictions about upcoming events. Such statements vary in the degree of certainty the author expresses toward the outcome: “Leonardo DiCaprio will win Best Actor” vs. “Leonardo DiCaprio may win” or “No way Leonardo wins!”. Can popular beliefs on social media predict who will win? To answer this question, we build a corpus of tweets annotated for veridicality on which we train a log-linear classifier that detects positive veridicality with high precision.[1] We then forecast uncertain outcomes using the *wisdom of crowds*, by aggregating users’ explicit predictions. Our method for forecasting winners is fully automated, relying only on a set of contenders as input. It requires no training data of past outcomes and outperforms sentiment and tweet volume baselines on a broad range of contest prediction tasks. We further demonstrate how our approach can be used to measure the reliability of individual accounts’ predictions and retrospectively identify surprise outcomes.

1 Introduction

In the digital era we live in, millions of people broadcast their thoughts and opinions online. These include predictions about upcoming events of yet unknown outcomes, such as the Oscars or election results. Such statements vary in the extent to which their authors intend to convey the event will happen. For instance, (a) in Table 1 strongly asserts the win of Natalie Portman over Meryl Streep, whereas (b) imbues the claim with uncertainty. In contrast, (c) does not say anything about the likelihood of Natalie Portman winning (although it clearly indicates the author would like her to win).

(a)	*Natalie Portman is gonna beat out Meryl Streep for best actress*
(b)	*La La Land doesn’t have lead actress and actor guaranteed. Natalie Portman will probably (and should) get best actress*
(c)	*Adored #LALALAND but it’s #NataliePortman who deserves the best actress #oscar #OscarNoms > superb acting*

Table 1: Examples of tweets expressing varying degrees of veridicality toward Natalie Portman winning an Oscar.

Prior work has made predictions about contests such as NFL games (Sinha et al., 2013) and elections using tweet volumes (Tumasjan et al., 2010) or sentiment analysis (O’Connor et al., 2010; Shi et al., 2012). Many such indirect signals have been shown useful for prediction, however their utility varies across domains. In this paper we explore whether the “wisdom of crowds” (Surowiecki, 2005), as measured by users’ explicit predictions, can predict outcomes of future events. We show how it is possible to accurately forecast winners, by aggregating many individual predictions that assert an outcome. Our approach requires no historical data about outcomes for training and can directly be adapted to a broad range of contests.

To extract users’ predictions from text, we present TwiVer, a system that classifies veridicality toward future contests with uncertain outcomes. Given a list of contenders competing in a contest (e.g., Academy Award for Best Actor), we use TwiVer to count how many tweets explicitly assert the win of each contender. We find that aggregating veridicality in this way provides

[1] The code and data can be found at `https://github.com/SandeshS/Twitter-Veridicality`

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1582–1591
Copenhagen, Denmark, September 7–11, 2017. ©2017 Association for Computational Linguistics

an accurate signal for predicting outcomes of future contests. Furthermore, TwiVer allows us to perform a number of novel qualitative analyses including retrospective detection of *surprise outcomes* that were not expected according to popular belief (Section 4.5). We also show how TwiVer can be used to measure the number of correct and incorrect predictions made by individual accounts. This provides an intuitive measurement of the reliability of an information source (Section 4.6).

2 Related Work

In this section we summarize related work on text-driven forecasting and computational models of veridicality.

Text-driven forecasting models (Smith, 2010) predict future response variables using text written in the present: e.g., forecasting films' box-office revenues using critics' reviews (Joshi et al., 2010), predicting citation counts of scientific articles (Yogatama et al., 2011) and success of literary works (Ashok et al., 2013), forecasting economic indicators using query logs (Choi and Varian, 2012), improving influenza forecasts using Twitter data (Paul et al., 2014), predicting betrayal in online strategy games (Niculae et al., 2015) and predicting changes to a knowledge-graph based on events mentioned in text (Konovalov et al., 2017). These methods typically require historical data for fitting model parameters, and may be sensitive to issues such as concept drift (Fung, 2014). In contrast, our approach does not rely on historical data for training; instead we forecast outcomes of future events by directly extracting users' explicit predictions from text.

Prior work has also demonstrated that user sentiment online directly correlates with various real-world time series, including polling data (O'Connor et al., 2010) and movie revenues (Mishne and Glance, 2006). In this paper, we empirically demonstrate that veridicality can often be more predictive than sentiment (Section 4.1).

Also related is prior work on *detecting veridicality* (de Marneffe et al., 2012; Søgaard et al., 2015) and sarcasm (González-Ibánez et al., 2011). Soni et al. (2014) investigate how journalists frame quoted content on Twitter using predicates such as *think*, *claim* or *admit*. In contrast, our system TwiVer, focuses on the author's belief toward a claim and direct predictions of future events as opposed to quoted content.

Our approach, which aggregates predictions extracted from user-generated text is related to prior work that leverages explicit, positive veridicality, statements to make inferences about users' demographics. For example, Coppersmith et al. (2014; 2015) exploit users' self-reported statements of diagnosis on Twitter.

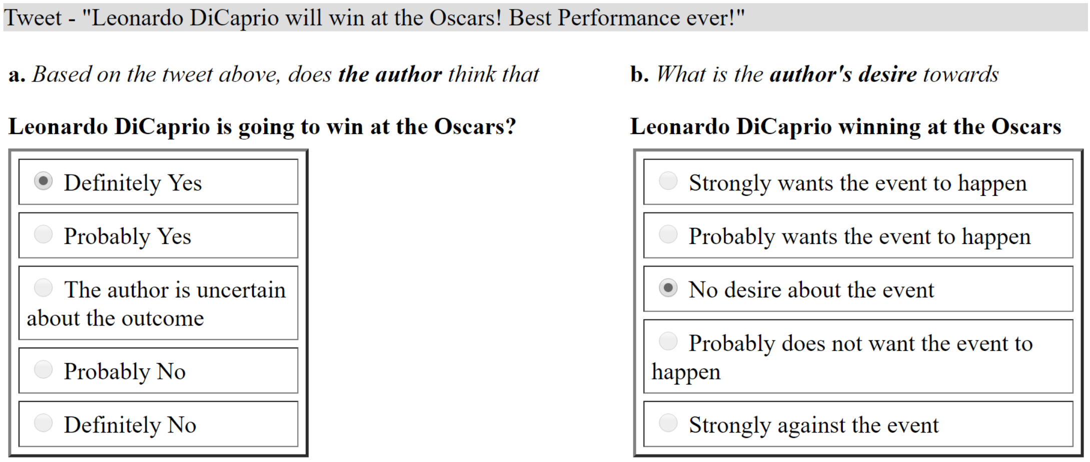
3 Measuring the Veridicality of Users' Predictions

The first step of our approach is to extract statements that make explicit predictions about unknown outcomes of future events. We focus specifically on *contests* which we define as events planned to occur on a specific date, where a number of *contenders* compete and a single *winner* is chosen. For example, Table 2 shows the contenders for Best Actor in 2016, highlighting the winner.

Actor	Movie
Leonardo DiCaprio	**The Revenant**
Bryan Cranston	Trumbo
Matt Damon	The Martian
Michael Fassbender	Steve Jobs
Eddie Redmayne	The Danish Girl

Table 2: Oscar nominations for Best Actor 2016.

To explore the accuracy of user predictions in social media, we gathered a corpus of tweets that mention events belonging to one of the 10 types listed in Table 4. Relevant messages were collected by formulating queries to the Twitter search interface that include the name of a contender for a given contest in conjunction with the keyword *win*. We restricted the time range of the queries to retrieve only messages written before the time of the contest to ensure that outcomes were unknown when the tweets were written. We include 10 days of data before the event for the presidential primaries and the final presidential elections, 7 days for the Oscars, Ballon d'Or and Indian general elections, and the period between the semi-finals and the finals for the sporting events. Table 3 shows several example queries to the Twitter search interface which were used to gather data. We automatically generated queries, using templates, for events scraped from various websites: 483 queries were generated for the presidential primaries based on events scraped from ballotpe-



Tweet - "Leonardo DiCaprio will win at the Oscars! Best Performance ever!"

a. *Based on the tweet above, does* ***the author*** *think that*

Leonardo DiCaprio is going to win at the Oscars?

- ● Definitely Yes
- ○ Probably Yes
- ○ The author is uncertain about the outcome
- ○ Probably No
- ○ Definitely No

b. *What is the* ***author's desire*** *towards*

Leonardo DiCaprio winning at the Oscars

- ○ Strongly wants the event to happen
- ○ Probably wants the event to happen
- ● No desire about the event
- ○ Probably does not want the event to happen
- ○ Strongly against the event

Figure 1: Example of one item to be annotated, as displayed to the Turkers.

dia[2] , 176 queries were generated for the Oscars,[3] 18 for Ballon d'Or,[4] 162 for the Eurovision contest,[5] 52 for Tennis Grand Slams,[6] 6 for the Rugby World Cup,[7] 18 for the Cricket World Cup,[8] 12 for the Football World Cup,[9] 76 for the 2016 US presidential elections,[10] and 68 queries for the 2014 Indian general elections.[11]

We added an event prefix (e.g., "Oscars" or the state for presidential primaries), a keyword ("win"), and the relevant date range for the event. For example, "Oscars Leonardo DiCaprio win since:2016-2-22 until:2016-2-28" would be the query generated for the first entry in Table 2.

Minnesota Rubio win since:2016-2-18 until:2016-3-1
Vermont Sanders win since:2016-2-18 until:2016-3-1
Oscars Sandra Bullock win since:2010-3-1 until:2010-3-7
Oscars Spotlight win since:2016-2-22 until:2016-2-28

Table 3: Examples of queries to extract tweets.

We restricted the data to English tweets only, as tagged by *langid.py* (Lui and Baldwin, 2012). Jaccard similarity was computed between messages to identify and remove duplicates.[12] We removed URLs and preserved only tweets that mention contenders in the text. This automatic post-processing left us with 57,711 tweets for all winners and 55,558 tweets for losers (contenders who did not win) across all events. Table 4 gives the data distribution across event categories.

Event	Number of tweets	
	Winners	Losers
2016 US Presidential primaries	20,347	17,873
Oscars (2009 – 2016)	1,498	872
Tennis Grand Slams (2011 – 2016)	10,785	19,745
Ballon d'Or Award (2010 – 2016)	3,492	3,285
Eurovision (2010 – 2016)	261	1,421
2016 US Presidential elections	9,679	3,966
2014 Indian general elections	920	736
Rugby World Cup (2010 – 2016)	272	379
Football World Cup (2010 – 2016)	8,129	5,489
Cricket World Cup (2010 – 2016)	2,328	1,792

Table 4: Number of tweets for each event category.

3.1 Mechanical Turk Annotation

We obtained veridicality annotations on a sample of the data using Amazon Mechanical Turk. For each tweet, we asked Turkers to judge veridicality toward a candidate winning as expressed in the tweet as well as the author's desire toward the event. For veridicality, we asked Turkers to rate whether the author believes the event will happen on a 1-5 scale ("Definitely Yes", "Probably Yes", "Uncertain about the outcome", "Probably No",

[2] https://ballotpedia.org/Main_Page

[3] https://en.wikipedia.org/wiki/Academy_Awards

[4] https://en.wikipedia.org/wiki/Ballon_d%27Or

[5] https://en.wikipedia.org/wiki/Eurovision_Song_Contest

[6] https://en.wikipedia.org/wiki/Grand_Slam_(tennis)

[7] https://en.wikipedia.org/wiki/Rugby_World_Cup

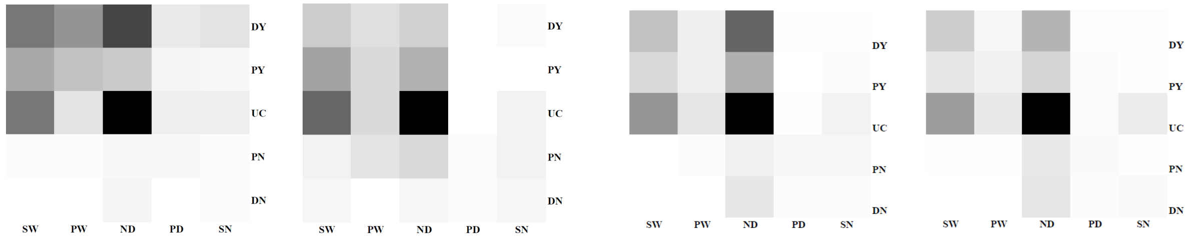
[8] https://en.wikipedia.org/wiki/Cricket_World_Cup

[9] https://en.wikipedia.org/wiki/FIFA_World_Cup

[10] https://en.wikipedia.org/wiki/United_States_presidential_election,_2016

[11] https://en.wikipedia.org/wiki/Indian_general_election,_2014

[12] A threshold of 0.7 was used.



(a) Oscar winners

(b) Oscar losers

(c) All events winners

(d) All events losers

Figure 2: Heatmaps showing annotation distributions for one of the events - the Oscars and all event types, separating winners from losers. Vertical labels indicate veridicality (DY "Definitely Yes", PY "Probably Yes", UC "Uncertain about the outcome", PN "Probably No" and DN "Definitely No"). Horizontal labels indicate desire (SW "Strongly wants the event to happen", PW "Probably wants the event to happen", ND "No desire about the event outcome", PD "Probably does not want the event to happen", SN "Strongly against the event happening"). More data in the upper left hand corner indicates there are more tweets with positive veridicality and desire.

"Definitely No"). We also added a question about the author's desire toward the event to make clear the difference between veridicality and desire. For example, "I really want Leonardo to win at the Oscars!" asserts the author's desire toward Leonardo winning, but remains agnostic about the likelihood of this outcome, whereas "Leonardo DiCaprio will win the Oscars" is predicting with confidence that the event will happen.

Figure 1 shows the annotation interface presented to Turkers. Each HIT contained 10 tweets to be annotated. We gathered annotations for $1,841$ tweets for winners and $1,702$ tweets for losers, giving us a total of $3,543$ tweets. We paid \$0.30 per HIT. The total cost for our dataset was \$1,000. Each tweet was annotated by 7 Turkers. We used MACE (Hovy et al., 2013) to resolve differences between annotators and produce a single gold label for each tweet.

Figures 2a and 2c show heatmaps of the distribution of annotations for the winners for the Oscars in addition to all categories. In both instances, most of the data is annotated with "Definitely Yes" and "Probably Yes" labels for veridicality. Figures 2b and 2d show that the distribution is more diverse for the losers. Such distributions indicate that the veridicality of crowds' statements could indeed be predictive of outcomes. We provide additional evidence for this hypothesis using automatic veridicality classification on larger datasets in §4.

3.2 Veridicality Classifier

The goal of our system, TwiVer, is to automate the annotation process by predicting how veridical a tweet is toward a candidate winning a contest: is the candidate deemed to be winning, or is the author uncertain? For the purpose of our experiments, we collapsed the five labels for veridicality into three: positive veridicality ("Definitely Yes" and "Probably Yes"), neutral ("Uncertain about the outcome") and negative veridicality ("Definitely No" and "Probably No").

We model the conditional distribution over a tweet's veridicality toward a candidate c winning a contest against a set of opponents, O, using a log-linear model:

$$P(y = v|c, \text{tweet}) \propto \exp\left(\theta_v \cdot f(c, O, \text{tweet})\right)$$

where v is the veridicality (positive, negative or neutral).

To extract features $f(c, O, \text{tweet})$, we first preprocessed tweets retrieved for a specific event to identify named entities, using (Ritter et al., 2011)'s Twitter NER system. Candidate (c) and opponent entities were identified in the tweet as follows:

- TARGET (t). A target is a named entity that matches a contender name from our queries.
- OPPONENT (O). For every event, along with the current TARGET entity, we also keep track of other contenders for the same event. If a named entity in the tweet matches with one of other contenders, it is labeled as opponent.
- ENTITY (e): Any named entity which does not match the list of contenders.

Figure 3 illustrates the named entity labeling for a tweet obtained from the query "Oscars Leonardo DiCaprio win since:2016-2-22 until:2016-2-28". Leonardo DiCaprio is the TARGET, while the named entity tag for Bryan Cranston, one of the

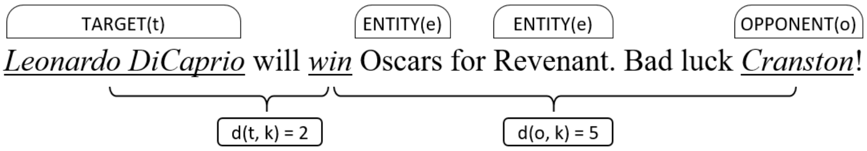

TARGET(t)
ENTITY(e)
ENTITY(e)
OPPONENT(o)
Leonardo DiCaprio will win Oscars for Revenant. Bad luck Cranston!
d(t, k) = 2
d(o, k) = 5

Figure 3: Illustration of the three named entity tags and distance features between entities and keyword *win* for a tweet retrieved by the query "Oscars Leonardo DiCaprio win since:2016-2-22 until:2016-2-28".

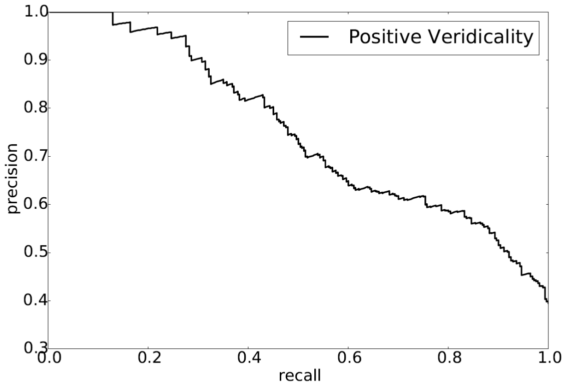


Figure 4: Precision/Recall curve showing TwiVer performance in identifying positive veridicality tweets in the test data.

	P	R	F1
– Context	47.7	**96.4**	63.8
– Distance	57.5	82.5	67.7
– Punctuation	53.4	88.2	66.6
– Dependency path	56.9	85.4	68.2
– Negated keyword	56.7	86.4	68.4
All features	**58.7**	83.1	**68.8**

Table 5: Feature ablation of the positive veridicality classifier by removing each group of features from the full set. The point of maximum F1 score is shown in each case.

losers for the Oscars, is re-tagged as OPPONENT. These tags provide information about the position of named entities relative to each other, which is used in the features.

3.3 Features

We use five feature templates: context words, distance between entities, presence of punctuation, dependency paths, and negated keyword.

Target and opponent contexts. For every TARGET (t) and OPPONENT ($o \in O$) entities in the tweet, we extract context words in a window of one to four words to the left and right of the TARGET ("Target context") and OPPONENT ("Opponent context"), e.g., *t will win*, *I'm going with* t, *o will win*.

Keyword context. For target and opponent entities, we also extract words between the entity and our specified keyword (k) (*win* in our case): *t predicted to* k, *o might* k.

Pair context. For the election type of events, in which two target entities are present (contender and state. e.g., *Clinton, Ohio*), we extract words between these two entities: e.g., t_1 *will win* t_2.

Distance to keyword. We also compute the distance of TARGET and OPPONENT entities to the keyword.

Punctuation. We introduce two binary features for the presence of exclamation marks and question marks in the tweet. We also have features which check whether a tweet ends with an exclamation mark, a question mark or a period. Punctuation, especially question marks, could indicate how certain authors are of their claims.

Dependency paths. We retrieve dependency paths between the two TARGET entities and between the TARGET and keyword (*win*) using the TweeboParser (Kong et al., 2014) after applying rules to normalize paths in the tree (e.g., "doesn't" → "does not").

Negated keyword. We check whether the keyword is negated (e.g., "not win", "never win"), using the normalized dependency paths.

We randomly divided the annotated tweets into a training set of 2,480 tweets, a development set of 354 tweets and a test set of 709 tweets. MAP parameters were fit using LBFGS-B (Zhu et al., 1997). Table 6 provides examples of high-weight features for positive and negative veridicality.

3.4 Evaluation

We evaluated TwiVer's precision and recall on our held-out test set of 709 tweets. Figure 4 shows the precision/recall curve for positive veridicality. By setting a threshold on the probability score to be greater than 0.64, we achieve a precision of

Positive Veridicality			Negative Veridicality		
Feature Type	Feature	Weight	Feature Type	Feature	Weight
Keyword context	TARGET *will* KEYWORD	0.41	Negated keyword	keyword is negated	0.47
Keyword dep. path	TARGET → *to* → KEYWORD	0.38	Keyword context	TARGET *won't* KEYWORD	0.41
Keyword dep. path	TARGET ← *is* → *going* → *to* → KEYWORD	0.29	Opponent context	OPPONENT *will win*	0.37
Target context	TARGET *is favored to win*	0.19	Keyword dep. path	TARGET ← *will* → *not* → KEYWORD	0.31
Keyword context	TARGET *are going to* KEYWORD	0.15	Distance to keyword	OPPONENT *closer to* KEYWORD	0.28
Target context	TARGET *predicted to win*	0.13	Target context	TARGET *may not win*	0.27
Pair context	TARGET1 *could win* TARGET2	0.13	Keyword dep. path	OPPONENT ← *will* → KEYWORD	0.23
Distance to keyword	TARGET *closer to* KEYWORD	0.11	Target context	TARGET *can't win*	0.18

Table 6: Some high-weight features for positive and negative veridicality.

Tweet	Gold	Predicted
The heart wants **Nadal** to win tomorrow but the mind points to a Djokovic win over 4 sets. Djokovic 7-5 4-6 7-5 6-4 **Nadal** for me.	negative	positive
Hopefully tomorrow Federer will win and beat that **Nadal** guy lol	neutral	negative
There is no doubt **India** have the tools required to win their second World Cup. Whether they do so will depend on ...	positive	neutral

Table 7: Some classification errors made by TwiVer. Contenders queried for are highlighted.

80.1% and a recall of 44.3% in identifying tweets expressing a positive veridicality toward a candidate winning a contest.

3.5 Performance on held-out event types

To assess the robustness of the veridicality classifier when applied to new types of events, we compared its performance when trained on all events vs. holding out one category for testing. Table 9 shows the comparison: the second and third columns give F1 score when training on all events vs. removing tweets related to the category we are testing on. In most cases we see a relatively modest drop in performance after holding out training data from the target event category, with the exception of elections. This suggests our approach can be applied to new event types without requiring in-domain training data for the veridicality classifier.

3.6 Error Analysis

Table 7 shows some examples which TwiVer incorrectly classifies. These errors indicate that even though shallow features and dependency paths do a decent job at predicting veridicality, deeper text understanding is needed for some cases. The opposition between "the heart ... the mind" in the first example is not trivial to capture. Paying attention to matrix clauses might be important too (as shown in the last tweet "There is no doubt ...").

4 Forecasting Contest Outcomes

We now have access to a classifier that can automatically detect positive veridicality predictions about a candidate winning a contest. This enables us to evaluate the accuracy of the crowd's wisdom by retrospectively comparing popular beliefs (as extracted and aggregated by TwiVer) against known outcomes of contests.

We will do this for each award category (Best Actor, Best Actress, Best Film and Best Director) in the Oscars from 2009 – 2016, for every state for both Republican and Democratic parties in the 2016 US primaries, for both the candidates in every state for the final 2016 US presidential elections, for every country in the finals of Eurovision song contest, for every contender for the Ballon d'Or award, for every party in every state for the 2014 Indian general elections, and for the contenders in the finals for all sporting events.

4.1 Prediction

A simple voting mechanism is used to predict contest outcomes: we collect tweets about each contender written before the date of the event,[13] and use TwiVer to measure the veridicality of users' predictions toward the events. Then, for each contender, we count the number of tweets that are labeled as positive with a confidence above 0.64, as well as the number of tweets with positive veridicality for all other contenders. Table 11 illustrates these counts for one contest, the Oscars Best Actress in 2014.

We then compute a simple prediction score, as follows:

$$\text{score} = (|T_c| + 1)/(|T_c| + |T_O| + 2) \quad (1)$$

[13] These are a different set of tweets than those TwiVer was trained on.

Event	Veridicality			Sentiment			Frequency			#predictions
	P	R	F1	P	R	F1	P	R	F1	
Oscars	**80.6**	80.6	**80.6**	52.9	87.1	63.5	54.7	**93.5**	69.0	151
Ballon d'Or	**100.0**	100.0	**100.0**	85.7	100.0	92.2	85.7	100.0	92.2	18
Eurovision	**83.3**	71.4	**76.8**	38.5	71.4	50.0	50.0	57.1	53.3	87
Tennis Grand Slam	50.0	100.0	66.6	50.0	100.0	66.6	50.0	100.0	66.6	52
Rugby World Cup	**100.0**	100.0	**100.0**	50.0	100.0	66.6	50.0	100.0	66.6	4
Cricket World Cup	**66.7**	85.7	**75.0**	58.3	**100.0**	73.6	58.3	**100.0**	73.6	14
Football World Cup	**71.4**	100.0	**83.3**	62.5	100.0	76.9	**71.4**	100.0	**83.3**	10
Presidential primaries	**66.0**	**88.0**	**75.4**	58.9	82.5	68.7	63.4	78.7	70.2	211
2016 US presidential elections	60.9	**100.0**	**75.6**	63.3	73.8	68.1	**69.0**	69.0	69.0	84
2014 Indian general elections	**95.8**	100.0	**97.8**	65.6	91.3	76.3	56.1	100.0	71.8	52

Table 8: Performance of Veridicality, Sentiment baseline, and Frequency baseline on all event categories (%).

Event	Train on all	Train without held-out event	$\|T_t\|$
Oscars	69.5	63.8	64
Ballon d'Or	54.6	46.6	61
Eurovision	65.7	63.2	48
Tennis Grand Slams	52.1	45.5	44
Rugby World Cup	56.5	58.1	44
Cricket World Cup	61.9	66.8	49
Football World Cup	76.0	67.5	56
Presidential primaries	59.8	48.1	117
2016 US presidential elections	52.0	52.3	54
Indian elections	60.3	39.0	44

Table 9: F1 scores for each event when training on all events vs. holding out that event from training. $|T_t|$ is the number of tweets of that event category present in the test dataset.

where $|T_c|$ is the set of tweets mentioning positive veridicality predictions toward candidate c, and $|T_O|$ is the set of all tweets predicting any opponent will win. For each contest, we simply predict as winner the contender whose score is highest.

4.2 Sentiment Baseline

We compare the performance of our approach against a state-of-the-art sentiment baseline (Mohammad et al., 2013). Prior work on social media analysis used sentiment to make predictions about real-world outcomes. For instance, O'Connor et al. (2010) correlated sentiment with public opinion polls and Tumasjan et al. (2010) use political sentiment to make predictions about outcomes in German elections.

We use a re-implementation of (Mohammad et al., 2013)'s system[14] to estimate sentiment for tweets in our corpus. We run the tweets obtained for every contender through the sentiment analysis system to obtain a count of positive labels. Sentiment scores are computed analogously to veridicality using Equation (1). For each contest, the contender with the highest sentiment prediction score is predicted as the winner.

4.3 Frequency Baseline

We also compare our approach against a simple frequency (tweet volume) baseline. For every contender, we compute the number of tweets that has been retrieved. Frequency scores are computed in the same way as for veridicality and sentiment using Equation (1). For every contest, the contender with the highest frequency score is selected to be the winner.

4.4 Results

Table 8 gives the precision, recall and max-F1 scores for veridicality, sentiment and volume-based forecasts on all the contests. The veridicality-based approach outperforms sentiment and volume-based approaches on 9 of the 10 events considered. For the Tennis Grand Slam, the three approaches perform poorly. The difference in performance for the veridicality approach is quite lower for the Tennis events than for the other events. It is well known however that winners of tennis tournaments are very hard to predict. The performance of the players in the last minutes

[14] `https://github.com/ntietz/tweetment`

		Veridicality			Sentiment		
		Contender		Score	Contender		Score
OSCARS		**Leonardo DiCaprio**		0.97	**Julianne Moore**		0.85
		Natalie Portman		0.92	Mickey Rourke		0.83
		Julianne Moore		0.91	**Leonardo DiCaprio (2016)**		0.82
		Daniel Day-Lewis		0.90	**Kate Winslet**		0.75
		Slumdog Millionaire		0.75	Leonardo DiCaprio (2014)		0.69
		Matthew McConaughey		0.74	**Slumdog Millionaire**		0.67
	!	The Revenant		0.73	**Danny Boyle**		0.67
		Argo		0.71	**Daniel Day-Lewis**		0.66
		Brie Larson		0.70	**Brie Larson**		0.65
		The Artist		0.67	George Miller		0.63
PRIMARIES		**Trump**	South Carolina	0.96	**Sanders**	West Virginia	0.96
		Clinton	Iowa	0.90	**Clinton**	North Carolina	0.93
		Trump	Massachusetts	0.88	**Trump**	North Carolina	0.91
		Trump	Tennessee	0.88	**Sanders**	Wyoming	0.90
		Sanders	Maine	0.87	**Sanders**	Oklahoma	0.89
		Sanders	Alaska	0.87	**Sanders**	Hawaii	0.86
	!	Trump	Maine	0.87	Sanders	Arizona	0.86
		Sanders	Wyoming	0.86	**Sanders**	Maine	0.85
		Trump	Louisiana	0.86	**Trump**	Delaware	0.84
	!	Clinton	Indiana	0.85	**Trump**	West Virginia	0.83

Table 10: Top 10 predictions of winners for Oscars and primaries based on veridicality and sentiment scores. Correct predictions are highlighted. "**!**" indicates a loss which wasn't expected.

Contender	$\|T_c\|$	$\|T_O\|$
Cate Blanchett	73	46
Amy Adams	6	113
Sandra Bullock	22	97
Judi Dench	2	117
Meryl Streep	16	103

Table 11: Positive veridicality tweet counts for the Best Actress category in 2014: $|T_c|$ is the count of positive veridicality tweets for the contender under consideration and $|T_O|$ is the count of positive veridicality tweets for the other contenders.

of the match are decisive, and even professionals have a difficult time predicting tennis winners.

Table 10 shows the 10 top predictions made by the veridicality and sentiment-based systems on two of the events we considered - the Oscars and the presidential primaries, highlighting correct predictions.

4.5 Surprise Outcomes

In addition to providing a general method for forecasting contest outcomes, our approach based on veridicality allows us to perform several novel analyses including retrospectively identifying surprise outcomes that were unexpected according to popular beliefs.

In Table 10, we see that the veridicality-based approach incorrectly predicts *The Revenant* as winning Best Film in 2016. This makes sense, because the film was widely expected to win at the time, according to popular belief. Numerous sources in the press,[15,16,17] qualify *The Revenant* not winning an Oscar as a big surprise.

Similarly, for the primaries, the two incorrect predictions made by the veridicality-based approach were surprise losses. News articles[18,19,20] indeed reported the loss of Maine for Trump and the loss of Indiana for Clinton as unexpected.

4.6 Assessing the Reliability of Accounts

Another nice feature of our approach based on veridicality is that it immediately provides an intuitive assessment on the reliability of individual Twitter accounts' predictions. For a given account, we can collect tweets about past contests, and extract those which exhibit positive veridicality toward the outcome, then simply count how often

[15] www.forbes.com/sites/zackomalleygreenburg/2016/02/29/spotlight-best-picture-oscar-is-surprise-of-the-night/#52f546c2721a

[16] www.vox.com/2016/2/26/11115788/revenant-best-picture

[17] www.mirror.co.uk/tv/tv-news/spotlight-wins-best-picture-2016-7460633

[18] http://patch.com/us/across-america/maine-republican-caucus-live-results-trump-favored-win-0

[19] http://www.huffingtonpost.com/entry/ted-cruz-upset-win-maine-republican-caucus_us_56db461ee4b0ffe6f8e9a865

[20] https://news.vice.com/article/bernie-sanders-wins-indiana-primary-in-surprise-upset-over-hillary-clinton

User account	Accuracy(%)	User account	Accuracy(%)
User 1	100 (out of 6)	twitreporting	100 (out of 3)
Cr7Prince4ever	100 (out of 6)	User 3	100 (out of 3)
goal_ghana	100 (out of 4)	Naijawhatsup	100 (out of 3)
User 2	100 (out of 4)	1Mrfutball	90 (out of 10)
breakingnewsnig	100 (out of 4)	User 4	77 (out of 13)

Table 12: List of users sorted by how accurate they were in their Ballon d'Or predictions.

the accounts were correct in their predictions.

As proof of concept, we retrieved within our dataset, the user names of accounts whose tweets about Ballon d'Or contests were classified as having positive veridicality. Table 12 gives accounts that made the largest number of correct predictions for Ballon d'Or awards between 2010 to 2016, sorted by users' prediction accuracy. Usernames of non-public figures are anonymized (as user 1, etc.) in the table. We did not extract more data for these users: we only look at the data we had already retrieved. Some users might not make predictions for all contests, which span 7 years.

Accounts like "goal_ghana", "breakingnewsnig" and "1Mrfutball", which are automatically identified by our analysis, are known to post tweets predominantly about soccer.

5 Conclusions

In this paper, we presented TwiVer, a veridicality classifier for tweets which is able to ascertain the degree of veridicality toward future contests. We showed that veridical statements on Twitter provide a strong predictive signal for winners on different types of events, and that our veridicality-based approach outperforms a sentiment and frequency baseline for predicting winners. Furthermore, our approach is able to retrospectively identify surprise outcomes. We also showed how our approach enables an intuitive yet novel method for evaluating the reliability of information sources.

Acknowledgments

We thank our anonymous reviewers for their valuable feedback. We also thank the Clippers group and Wei Xu at The Ohio State University for their useful suggestions. This material is based upon work supported by the National Science Foundation under Grants No. IIS-1464128 to Alan Ritter and IIS-1464252 to Marie-Catherine de Marneffe. Alan Ritter is supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center in addition to the Office of the Director of National Intelligence (ODNI) and the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory (AFRL) contract number FA8750-16-C-0114. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, AFRL, NSF, or the U.S. Government.

References

Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. 2013. Success with style: Using writing style to predict the success of novels. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Hyunyoung Choi and Hal Varian. 2012. Predicting the present with Google Trends. *Economic Record*, 88(1):2–9.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. North American Chapter of the Association for Computational Linguistics.

Glen Coppersmith, Craig Harman, and Mark Dredze. 2014. Measuring post traumatic stress disorder in Twitter. In *International Conference on Weblogs and Social Media*.

Kaiser Fung. 2014. Google flu trends failure shows good data > big data. In *Harvard Business Review/HBR Blog Network[Online]*.

Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in Twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of NAACL-HLT*.

Mahesh Joshi, Dipanjan Das, Kevin Gimpel, and Noah A Smith. 2010. Movie reviews and revenues:

An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proceedings of EMNLP*.

Alexander Konovalov, Benjamin Strauss, Alan Ritter, and Brendan O'Connor. 2017. Learning to extract events from knowledge base revisions. In *Proceedings of WWW*.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the 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.

Gilad Mishne and Natalie S Glance. 2006. Predicting movie sales from blogger sentiment. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises*.

Vlad Niculae, Srijan Kumar, Jordan Boyd-Graber, and Cristian Danescu-Niculescu-Mizil. 2015. Linguistic harbingers of betrayal: A case study on an online strategy game. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

Michael J Paul, Mark Dredze, and David Broniatowski. 2014. Twitter improves influenza forecasting. *PLOS Currents Outbreaks*, 6.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Lei Shi, Neeraj Agarwal, Ankur Agarwal, Rahul Garg, and Jacob Spoelstra. 2012. Predicting US primary elections with Twitter. In *Social Network and Social Media Analysis: Methods, Models and Applications, NIPS*.

Shiladitya Sinha, Chris Dyer, Kevin Gimpel, and Noah A. Smith. 2013. Predicting the NFL using Twitter. In *Proceedings of ECML/PKDD Workshop on Machine Learning and Data Mining for Sports Analytics*.

Noah A. Smith. 2010. *Text-driven forecasting*. http://www.cs.cmu.edu/ nasmith/papers/smith.whitepaper10.pdf.

Anders Søgaard, Barbara Plank, and Hector Martinez Alonso. 2015. Using frame semantics for knowledge extraction from Twitter. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

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*.

James Surowiecki. 2005. *The wisdom of crowds*. Anchor Books, New York, NY.

Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

Dani Yogatama, Michael Heilman, Brendan O'Connor, Chris Dyer, Bryan R Routledge, and Noah A Smith. 2011. Predicting a scientific community's response to an article. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. 1997. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*.