# Recognizing Counterfactual Thinking in Social Media Texts

**Youngseo Son**[†], **Anneke Buffone**[§], **Anthony Janocko**[§], **Allegra Larche**[§], **Joseph Raso**[§], **Kevin Zembroski**[§], **H Andrew Schwartz**[†], **Lyle Ungar**[§]
[†]Stony Brook University, [§]University of Pennsylvania
yson@cs.stonybrook.edu, ungar@cis.upenn.edu

## Abstract

Counterfactual statements, describing events that did *not* occur and their consequents, have been studied in areas including problem-solving, affect management, and behavior regulation. People with more counterfactual thinking tend to perceive life events as more personally meaningful. Nevertheless, counterfactuals have not been studied in computational linguistics. We create a counterfactual tweet dataset and explore approaches for detecting counterfactuals using rule-based and supervised statistical approaches. A combined rule-based and statistical approach yielded the best results (F1 = 0.77) outperforming either approach used alone.

## 1 Introduction

Counterfactuals describe events that did not occur, and what would have happened (or not happened), had the event occurred (e.g., "If I hadn't broken my arm, I never would have met her."). More precisely, counterfactual conditionals have the form "If it had been the case that A (or not A), it would have been the case that B (or not B)."

Counterfactuals have been studied in many different domains. Logicians and philosophers focus on literally logical relations between the antecedent and consequent of counterfactual forms and the outcomes (Goodman, 1947). In contrast, political scientists usually conduct counterfactual thought experiments for hypothetical tests on historical events, policies, or other aspects of a society and assess them (Tetlock, 1996).

Counterfactual thoughts are defined, especially in psychology, as mental representations of alternatives to past events, actions, or states. Their use has been explored for correlations with many different demographics (age, gender) and psychological variables (depression, religiosity) (Kray et al., 2010; Markman and Miller, 2006).

Counterfactual thinking has been linked to perceiving life events as more meaningful, fated, and even as influenced by the divine (Kray et al., 2010; Buffone et al., 2016), as well as with problem-solving, because imagining alternate outcomes can easily bring to mind the steps needed for improvement (Epstude and Roese, 2008; Roese, 1994). It has also been shown to be associated with affect management, particularly when imagining realities that are worse than what actually happened (Epstude and Roese, 2008; Roese, 1994)

Despite the extensive research on counterfactual *thinking*, counterfactual *language forms* have not been studied in computational linguistics. Language-based models to recognize counterfactual thinking in social media would potentially allow for psychological analysis on users based on their everyday language, avoiding the high expense of capturing counterfactual thinking at a large scale using traditional psychological assessments.

Therefore, in this paper, we build a language-based model to recognize counterfactual forms in social media texts of Twitter and Facebook. There are many challenges for this task. First, counterfactual statements have a low base rate; we found only 2% of status updates on Facebook and 1% of tweets contain counterfactual statements. Secondly, counterfactual statements can take on many forms in natural language.[1] For example, they may or may not use explicit if- or then- clauses (e.g, consider "If I had not met him then I would be better off" versus "I wish I had not met him").

[1]Simply looking for words like 'if' fails to produce useful results; only 2 percent of sentences of tweets containing 'if' are counterfactuals.

The low base rate and high variability of natural language counterfactuals in social media texts make them difficult to recognize using simple linguistic or statistical features. We address these challenges by using a combined rule-based and statistical approach. Key to our success is defining seven sub-types of counterfactuals, allowing better coverage of rarer sub-types.

## 2 Related Work

Identifying counterfactuals is in many ways similar to identifying discourse relations. In terms of relation classification, the counterfactual conditionals can be viewed as a subset of Condition type of Contingency class in the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) or the Condition relation of Rhetorical Structure Theory (RST) (Mann and Thompson, 1987). Also, like all discourse relations in the PDTB, counterfactuals have implicit and explicit forms, and so cannot by uniquely identified by the presence of specific words.

There have been many researchers who have tried end-to-end discourse relation parsing with the PDTB and RST (Biran and McKeown, 2015; Lin et al., 2014; Ji and Eisenstein, 2014). Many of them used dependency parsing or constituency parsing for argument detection or elementary discourse unit (EDU) segmentation to infer the relation between them. However, the short lengths and poor quality of parses of social media texts make dependency constituents unreliable.[2] For example, posters frequently drop the subject of a sentence.

Other work mostly focuses on relation classification with an assumption that arguments of the given relations are already identified (Park and Cardie, 2012; Pitler et al., 2009). They explore various learning algorithms and types of features in the given arguments of discourse relations. Then, they report which combinations give the best performance of each discourse relation.

Our work, while possible to view as a task in discourse relation classification, focuses on critical features of counterfactuals rather than on accurate demarcation of each argument of the relation. Most downstream applications, such as psychological studies, require knowing the presence/absence of counterfactuals rather than their exact extent.

## 3 Method

We use a combination of a rule-based approach and a supervised classifier to capture counterfactual statements from Twitter.

### 3.1 Data Set

No existing corpus of counterfactual statements was available, so we collected our own data set, starting from a random set of tweets from May 2014 and July 2014. As noted previously, couterfactual statements are rare, so we first limited the random tweet set to 1,637 containing keywords[3] that can signal counterfactuals (Train and Test row from Table 1). Keywords were in part based on prior literature on spontaneous counterfactual generation, such as should have, could have, at least, if only, or next time (Sanna and Turley, 1996). We identified further counterfactual forms (e.g. wish) based on visual inspection of the data. Next we used the overall list of keywords to draw samples of 500 tweets for further visual inspection. Words or phrases which had an unreasonably high false positive rate for containing counterfactuals were eliminated. Well-trained annotators then manually labeled each of the 1,637 tweets for counterfactuals with a 9% postive rate, results in 153 counterfactuals and 1,484 negative samples. A random set of 500 of these instances were used in training and the rest were reserved for testing.

To build out our training set to capture examples of all forms of counterfactuals, we added a *train supplement* from random tweets from 2012 – at least thirty tweets from each of seven counterfactual forms we defined for our statistical model using the regular expressions[4] with brown-clusters and the tweet PTB tagging model (described next). With this process, we enabled the model to be less biased towards only the samples with the counterfactual cue phrases used for data collection. Additionally, the model learned syntactically different forms of counterfactuals identified in prior work. To evaluate counterfactual form annotation, inter-annotator agreement was established on

[2] In our preliminary experiments on our causality tweet dataset ($\kappa = 0.61$), Lin's parser (Lin et al., 2014) obtained 0.45 F1 while a linear support vector machine (SVM) with n-gram obtained 0.58 F1 for causality detection.

[3] e.g., 'should', 'shulda'; full list available in our supplementary data.

[4] The regular expression table is included in our supplementary data.

| Dataset | CF | Non-CF | Total |
|---|---|---|---|
| Train | 49 | 451 | 500 |
| + Supplement | 768 | 498 | 1,266 |
| Test | 104 | 1,033 | 1,137 |

Table 1: Data Collection. 'CF' is counterfactual and 'Non-CF' is non-counterfactual

1,637 tweets with a second rater with achieving $\kappa = 0.774$ and human annotation F1 0.79.

### 3.2 Classification

We first use a rule-based model to capture counterfactual patterns from social media texts. We then use a statistical model (Linear SVM) to increase precision by identifying tricky false positives with forms similar to counterfactuals (e.g., "wish you the best").

**Rule-based Classification.** Our rule-based approach is based on seven forms of counterfactuals (Table 2). Central to our method is our theorizing, based on reading the literature, especially (Kray et al., 2010) and examining many counterfactual examples, that counterfactuals come in seven different forms, shown with examples in (Table 2). First, we remove sentences ending in question marks predicted as 'end of sentence' by the tweet part-of-speech (POS) tagger (Gimpel et al., 2011). We then use pattern matching with regular expression using a combination of cue phrases (bold), POS tags, and word clusters. The word clusters, based on a set of Twitter Brown clusters[5] are used to capture the numerous variations of words in social media texts (e.g., 'shuldve' for 'should have'). This approach requires matching both the token and its part-of-speech, since the POS tag of each token is important for counterfactual form.

The rule-based approach is also useful in that it allows us to detect the arguments for counterfactual relations; conditional statement and consequent statement from *Conjunctive Normal/ Converse* form and *Verb Inversion* form, one counterfactual statement from *Wish Verb* and *Could / Would / Should have*. We customized Biran's demarcation methods using the first verb phrase or the connective as a boundary to capture the more informative argument of the statement: For one argument detection, we demarcate from the cue phrase (e.g., would have) to the end of sentence. For two arguments, we demarcate from conditional word (e.g., if, unless) to the end of statement or before the start of the second verb phrase.

**Part of Speech Tagging** We use the Penn Treebank (PTB)-style Tweet POS tags[6] instead of Tweet POS tags (Gimpel et al., 2011) as it contains more fine-grained categories and yields higher accuracy of pattern matching. For instance, Tweet POS tags do not differentiate modal verbs, past tense verbs, and other types of verbs, but categorize all of them as 'V'. However, in many forms of counterfactuals, the distinction between modal verbs and past particles from other types of verbs are critical (e.g., in *Should / Could / Would Have* forms). Finally, we conduct a postprocessing on the Tweet POS parsing results for the more accurate prediction. First, we delete RT tags along with the token since it is not informative for our task. Then, we convert 'USR' to nouns because the word token tagged as 'user' usually plays the role of a common noun from the discourse relation perspective. Additionally, in order to enhance the POS tagging, we use the brown clusters to tag empirical variations of modal verbs as 'MD' and we define 'CCJ', a new tag to distinguish conditional conjunctions (i.e. Brown clusters for 'if') from other types of conjunctions.

**Statistical Modeling.** Each counterfactual form has a different number of arguments for the relation, and different types of features that cause the most errors. Therefore, we analyze the errors of each form separately and use different approaches expected to ensure the best performance.

If a tweet matches rules for counterfactual forms 1, 2, 3, 4, or 5, it is further classified using a statistical model trained with features of sequential words (n-gram) and POS tags of demarcated arguments and the whole sentence.

A statistical model is expected to capture some implicit relations between arguments as well as lexical and part-of-speech patterns, but may also hurt performance in situations where the rule-based approach achieves high precision. Therefore, we applied statistical approaches to counterfactual forms which cannot be easily differentiated by their superficial patterns. These forms were selected by both theoretical and empirical analysis; we discuss these forms further in our evaluation section.

[5] http://www.cs.cmu.edu/~ark/TweetNLP/clusters/50mpaths2

[6] http://www.cs.cmu.edu/~ark/TweetNLP/model.ritter_ptb_alldata_fixed.20130723

| Counterfactual Form | Example |
|---|---|
| 1. Wish Verb | I **wish I had** been richer |
| 2. Conjunctive Normal | **If** everyone **put** differences aside and get along, everything would be so much enjoyable |
| 3. Conjunctive Converse | I **would** be stronger, if I **had lifted** weights |
| 4. Modal Normal | They **should of shown** this guy gettin shot, that **woulda been** TV gold. |
| 5. Verb Inversion | **Had I left** the event early, I **would not have met** John |
| 6. Should Have | I **should have** joined the event early |
| 7. (Would / Could) Have | I **would have been happier** without John |

Table 2: Counterfactual Forms

| Performance | **CF Parser** | **Rules only** | **SVM** |
|---|---|---|---|
| Precision | 0.7131 | 0.5864 | 0.2381 |
| Recall | 0.8365 | 0.9134 | 0.9135 |
| F1 | **0.7699** | 0.7143 | 0.3777 |

Table 3: Performance of Classifiers

| **Process** | F1 | Precision | Recall |
|---|---|---|---|
| Whole Pipeline | 0.7699 | 0.7131 | 0.8365 |
| - Args | 0.7595 | 0.6767 | 0.8653 |
| - PTB | 0.7456 | 0.6854 | 0.8173 |
| - Form 1 | 0.7447 | 0.6592 | 0.8557 |
| - Form 2,3,4,5 | 0.7352 | 0.6241 | 0.8942 |

Table 4: Ablation Test for Each Process

## 4 Evaluation

As discussed, counterfactuals are not easily identified by rules or specific words. Given their low base rate and multiplicity of forms, traditional machine learning approaches trained on a random tweet sample tend to label all tweets as the most frequent class (non-counterfactual). Use of a counterfactual-enriched training set increases the performance, but still gives a low F1 on the imbalanced test set.

Thus, in order to make the classifier robust to the imbalanced dataset, we designed a rule-based model with counterfactual forms, which resulted in significantly higher F1 than statistical model. Moreover, the rule-based model captures positive samples of all possible forms which might not exist in the training set. A combined approach gives the best result. As Table 3 shows, our whole pipeline ('CF Parser' in Table 3) obtained the best overall performance with the combination of both approaches.

For *Wish Verb* form prediction gets a big performance boost from the statistical model because of highly frequent false positives which have counterfactual-like forms such as birthday wishes or new year's day wishes. Among samples classified as *Wish Verb* form the counterfactual prediction F1 increased from 0.82 to 0.90 after the final prediction by the statistical model.

Finally, we conducted an ablation test to analyze how each process of the pipeline affects the overall performance of the classifier (Table 4). The argument detection was less effective (F1 0.01 drop) than we expected due to the relatively simple and concise structure of tweets in general (Args in Table 4).

Using only n-grams as features for the statistical model without PTB-style Tweet POS tags gives a relatively large drop (0.02) from F1. From the grammatical perspective, n-grams are less informative than POS tags for counterfactuals especially considering that there are so many variations of each word token in social media (e.g., 'clda', 'coulda', and 'couldve' for 'could have').

We examined how the statistical model affected the final performance of each counterfactual form. The model we used for filtering out frequent false positives (e.g., birth day wishes) of *Wish Verb* form caused 0.03 F1 drop when it is removed. Also, the models trained with two-argument-relation forms (*Conjunctive Normal / Converse*, *Modal Normal*, and *Verb Inversion*) caused 0.04 F1 drop when they are removed from the pipeline, since the classifier cannot use subtle relations between arguments for its counterfactual prediction.

## 5 Conclusion

This is the first work to identify counterfactuals in social media, a task we hope more people will address. Our best results came from combining rule-based methods that exploit a theory of the different forms of counterfactual with focused statistical methods for reclassification of challenging forms. Our counterfactual predictor can now be applied to large collections of tweets and Facebook posts from people of known education, religiosity, political orientation, well-being, and other attributes of interest to psychologists and political scientists, allowing further study of their theories of counterfactual use.

## Acknowledgments

This project/publication was made possible, in part, through the support of a grant from Templeton Religion Trust, TRT-0048. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the Templeton Religion Trust.

## References

Or Biran and Kathleen McKeown. 2015. Pdtb discourse parsing as a tagging task: The two taggers approach. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. pages 96–104.

Anneke Buffone, Shira Gabriel, and Michael Poulin. 2016. There but for the grace of god counterfactuals influence religious belief and images of the divine. *Social Psychological and Personality Science* 7(3):256–263.

Kai Epstude and Neal J Roese. 2008. The functional theory of counterfactual thinking. *Personality and Social Psychology Review* 12(2):168–192.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, pages 42–47.

Nelson Goodman. 1947. The problem of counterfactual conditionals. *The Journal of Philosophy* 44(5):113–128.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *ACL (1)*. pages 13–24.

Laura J Kray, Linda G George, Katie A Liljenquist, Adam D Galinsky, Philip E Tetlock, and Neal J Roese. 2010. From what might have been to what must have been: counterfactual thinking creates meaning. *Journal of personality and social psychology* 98(1):106.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering* 20(02):151–184.

William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute.

Keith D Markman and Audrey K Miller. 2006. Depression, control, and counterfactual thinking: Functional for whom? *Journal of Social and Clinical Psychology* 25(2):210–227.

Joonsuk Park and Claire Cardie. 2012. Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, pages 108–112.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *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 2-Volume 2*. Association for Computational Linguistics, pages 683–691.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *In Proceedings of LREC*.

Neal J Roese. 1994. The functional basis of counterfactual thinking. *Journal of personality and Social Psychology* 66(5):805.

Lawrence J Sanna and Kandi Jo Turley. 1996. Antecedents to spontaneous counterfactual thinking: Effects of expectancy violation and outcome valence. *Personality and Social Psychology Bulletin* 22(9):906–919.

Philip E Tetlock. 1996. *Counterfactual thought experiments in world politics: Logical, methodological, and psychological perspectives*. Princeton University Press.