Counterfactuals in the Language of Social Media: A Natural Language Processing Project in Conjunction with the World Well Being Project

August 28, 2015 to April 25, 2016

Anthony Janocko, Allegra Larche, Joseph Raso, Kevin Zembroski

Advisor: Lyle Ungar

World Well Being Project Liaison: Anneke Buffone

University of Pennsylvania School of Engineering and Applied Science
Computer and Information Science Senior Design Project Report
Abstract

Certain aspects of natural language can give clues to a person’s personality and other traits. One such aspect is the use of counterfactual expressions. Counterfactuals are statements that examine how a hypothetical change in a past experience could have affected the outcome of that experience. Counterfactuals have been shown to bring meaning to people’s lives, alter behavior related to planning, and achieve affect and emotional management. Most counterfactuals are in the form of a conditional statement such as “if… then…” and contain modal verbs. Hypothetical statements regarding future events are not considered counterfactuals. After investigating patterns in counterfactual usage, we created a natural language processing model that tags tweets as either counterfactual or not. We predict that in United States counties, tweet counterfactual usage as a percentage of total tweets will be directly related to life satisfaction in a given county. Our model achieves 90.24% accuracy in identifying counterfactuals from our test set. The experiment determines a tweet’s location, updates a total tweet count, and updates a counterfactual count using the model for each county. The percentage counterfactual usage is used in a linear regression analysis to determine the validity of our hypothesis. Finally, a web application visualizes counterfactual usage by county and uses the Twitter API to allow users to set a time period and view counterfactuals in a given location. The application also allows users to view the counterfactuals in their personal profile.
Introduction

Previous research surrounding counterfactual thinking has analyzed how counterfactual usage is linked to certain aspects of individual’s personalities. For example, in 2010 research showed that counterfactual thinking surrounding certain events in subjects’ lives led them to assign more meaning to the event they analyzed. It has also been proven that people who employ counterfactual thinking are better at working in teams, and that counterfactual thinking is noticeably deficient in people with schizophrenia. Because of the significance of counterfactual thinking in identifying aspects of one’s personality, we decided that it would be useful to build a model which could correctly identify counterfactual usage on a person’s Twitter profile. We chose to use Twitter data because of the brevity of the messages to be analyzed and the tendency for people to talk about themselves via this platform.

After building the model for finding counterfactual statements on Twitter, our team decided that it would be interesting to see if counterfactual thinking within a community could be linked to any socio-economic statistics about that community. We hypothesized that counterfactual usage would be linked to a community’s overall life satisfaction, since at an individual level it was shown to bring meaning to life. Our results were largely inconclusive, but we present several other applications for our model later in this report.

Definitions and Forms

Conditional Conjunctions

A conditional conjunction is a conjunction that describes a condition. The use of a conditional conjunction implies that one clause of a sentence is dependent on another clause being possible. Thus, conditional conjunctions are used to describe hypothetical situations and to explain why something has happened, will happen or is currently happening at that moment in time. Conditional conjunctions include if, unless, since, as a consequence of, provided that, assuming that etc.

Examples:
1. If you brush your teeth every day, you probably won’t get cavities.
2. If you want me to, I can reorganize your desk so you have more room to work.

1 Kray, Laura J.; George, Linda G.; Liljenquist, Katie A.; Galinsky, Adam D.; Tetlock, Philip E.; Roese, Neal J. (2010). From what might have been to what must have been: Counterfactual thinking creates meaning. Journal of Personality and Social Psychology, Vol 98(1), 106-118.
3 Galinsky, A. D., & Kray, L. J. (2004). From thinking about what might have been to sharing what we know: The effects of counterfactual mind-sets on information sharing in groups. Journal of Experimental Social Psychology, 40, 606-618.
3. Since I'll be out of town, I'm looking for someone to feed my cat.

Modal Verbs

Modal verbs are used to express probability, ability, obligation, possibility, permission and requests. Common modal verbs include: can, could, may, might, will, would, must, shall, should and ought to. ⁵

Examples:
1. She can speak six languages.
2. He could have missed the train.
3. Children must do their homework.

Subjunctive Mood

The subjunctive mood refers to a hypothetical state or a state inconsistent with reality. For example, wishes, desires, imaginary situations and suggestions can be expressed using the subjunctive mood. In present day English, this verb mood has become increasingly rare as speakers opt to use conditional forms such as could or would. When considering counterfactuals, it is important to recognize this form because statements that employ subjunctive mood are often counterfactuals. ⁶

Examples:
1. If I were a butterfly, I would have wings.
2. I wish I had gone to the show.
3. A car might crash into his house if he were to build it near the Interstate.

Counterfactuals

We redefined the counterfactual by combining our understanding with various dictionary definitions. We also looked into defining further classifications of counterfactuals that may be useful in other research.

Counterfactual: a statement which looks at how a hypothetical change in a past experience could have affected the outcome of that experience.
Upward/Downward Counterfactual: a counterfactual in which the hypothetical reality is better/worse (for the person making the statement) than what actually occurred.
Additive/Subtractive Counterfactual: a counterfactual statement is additive if the antecedent contains something the person should have done. A subtractive is one where the antecedent contains something they should not have done.
Self/Other Counterfactual: a counterfactual statement relates to the the person making the statement if they had any role in the experience or outcome, or if they were affected in any way by the experience or outcome.
Antecedent: is the past hypothetical experience within a counterfactual.

**Consequence**: is the potential outcome within a counterfactual.

**Identified Forms**

During our research, we identified and categorized the following listed counterfactual forms. These forms are further split into an antecedent and consequence. Each of these forms employs conditional conjunctions, modal verbs, or the subjunctive mood. Some of these examples were taken from Twitter.

1. **Conjunctive Normal**: Antecedent -> Consequence
   The antecedent consists of a *conditional conjunction* followed by a *past tense subjunctive verb* or *past modal verb*.
   The consequence contains a *past or present tense modal verb*.
   a. If everyone *could* just put differences aside and get along, everything *would be* so much more enjoyable
   b. @tmorello *if* Obama was Bush you guys *would be screaming* a lot more. And you'd be right. No comparison needed to anyone. But it's Moyers so..

2. **Conjunctive Converse**: Consequence -> Antecedent
   The antecedent consists of a *modal verb* and *past or present tense verb*.
   The consequence consists of a *conditional conjunction* followed by a *past tense subjunctive verb* or *past tense modal*.
   a. I *would be* stronger, *if* I had lifted weights.

3. **Modal Normal**: Antecedent -> Consequence
   The antecedent consists of a *modal verb* and *past participle verb*.
   The consequence consists of a *past/present tense modal verb*.
   a. They *should of shown* this guy gettin shot, *that woulda been* TV gold

4. **Wish/Should Implied**: Antecedent ->
   The antecedent is the independent clause following 'wish' or 'should'.
   The consequence is implied and can be paraphrased as “would be better off.”
   a. I *wish I had* longer hair.
   b. I *should have gone* to the store.

5. **Verb Inversion**
   The antecedent uses the subjunctive mood by inverting the verbs 'had' and 'were' to create a hypothetical conditional statement along with a *past tense verb*.
   The consequence consists of a *modal verb* and *past or present tense verb*.
   Key words: had, were
   a. *Had I left* the event early, *I would not have met* John.
   b. *I could have fallen*, *had I not payed* attention.
6. **Modal Propositional:** Consequence -> Antecedent

The antecedent consists of a *past/present modal verb*. The consequence consists of a *prepositional phrase* (only certain types).

- Key words: with/without
- a. Thank you for sharing the Hub love, Omar! LA would be a lesser city without your work for the #lariver&#8211;

### Counterfactual Model

Our main focus for our model was to correctly classify counterfactuals within tweets. Our initial idea was to use some a machine learning model by looking at word use and placement within a sentence or tweet. However, as we thought more about how a counterfactual is formed within the English language, we soon realized a simpler and more effective approach. The subjunctive mood is a way of expressing an idea that is hypothetical. Within the English language it is expressed not as a verb tense, but with various key words. Counterfactuals fall within a subset of the subjunctive mood in the past tense. As we saw the patterns for how counterfactuals were created, we designed a new approach using a combination of a part of speech tagger and regular expressions.

The method works as follows. A tweet is tokenized using the Carnegie Mellon twokenizer package. Each token is then fed through the Python NLTK perceptron tagger. This was the most accessible tagger that gave us the important feature of verb tenses. The tagged tweet is then formatted so that each token is followed by its part of speech. A post-processing step tags some of the most common misspellings of important keywords like modal verbs. The post-processing also creates a new tag for conditional conjunctions such as ‘if’. We took some leniency with the subjunctive mood because of the improper usage on Twitter. In the final step, the tagged tweet is run through a series of regular expressions that correspond to the six different identified forms for counterfactuals excluding questions. The tweet is then binarily classified.

With this model, we used a training set of 1000 tweets that were filtered by a modal verbs in order to tweak our regular expressions. We were able to achieve a positive predictive value or precision of 63.0% and a sensitivity of 54.31% giving an accuracy of 91.0% on the filtered training set. We tested the model on a testing set of size 652 with slightly different keywords. This gave a precision of 45.31%, sensitivity of 64.44%, and accuracy of 92.18%.

### Model Improvements

The main issue with classification was the accuracy of the part of speech tagger. This was due to the improper formatting of sentences on Twitter, many common misspellings of
words (eg. shud for should), and contractions (eg. I'd, we'd). Many verbs tenses were ambiguous to the reader or the tagger gave present tense verb tags for past tense verb usage. A few past tense verbs were labeled as adjectives as well. This was an important distinction for counterfactual classification because of the necessity of a past tense verb. A tagger trained on Twitter data may offer a more robust way of correct verb form tagging.

Another issue that arose was the lack of a tag for conditional conjunctions. A few conditional conjunctions include: if, provided (that), supposing, imagine, and as long as. These are important words and phrases that are necessary for forming counterfactuals. Since these phrases were not included in the tagger, our post-processing step attempts to catch some of these words and tag them.

A more robust tagger along with a combination of a machine learning and regular expression model could provide the correct framework for identifying hypothetical and subjunctive mood. These classifications could then be further classified as counterfactuals or other hypothetical forms.

Web Application

After building our model, we wanted a way for users to interact with our model and data. We accomplished this by building a web application using the angularJS frontend framework and an express server. We also employed the Google Material Design principles when developing the application. Given that we separated our experiments into two distinct categories, experiments on the county-wide level and experiments on the personal level, it was natural to split our main application into two smaller applications.

Our first application focused on counterfactual usage at the personal level. We first hook up to the Twitter API in order to fetch a user’s recent tweets. From there we run these tweets through our classification scripts and display all of the tweets that we have identified as counterfactual. We then take their percent counterfactual usage and compare it to the data gathered by the World Well-Being Project to infer this user’s personality traits.

The second part of the application focused on counterfactual usage at the county level. We built an interface that allowed users to select a specific county and time frame and visually display all of the counterfactual tweets from that period. Based on the percent counterfactual usage in the given county, we can take the regression models that we created to infer various socioeconomic factors such as divorce rate and violent crime rate for this county. One of the setbacks we
faced while developing this portion of the application was the Twitter API search limitations. Our original plan was to pull tweets directly from Twitter, however we could only search up to 6 days in the past and couldn’t pull a sufficient number of tweets at one time. To remedy this problem, we ended up using tweets stored in the WWBP databases.

**Ethical/ Privacy Concerns**

Working with Twitter data naturally leads to some privacy concerns. We mitigated these concerns by only using publically available data when performing our statistical experiments. Additionally, the web app required that a user provide authorization before we pulled any tweets from their timeline for analyzing.

**Statistics Experiments**

**Methods**

In order to run correlation tests between county level socioeconomic factors and counterfactual usage, we used our counterfactual tagging model to identify 20 million tweets from the WWBP random twitter 2012 database as counterfactual or not. During the tagging, we determined each tweet’s county location based on the tweet’s coordinates. Next, we calculated a percentage counterfactual usage for each county by dividing the total number of counterfactuals by the total tweet count. In order to avoid outliers from counties with very few tweets, we decided to only include counties with at least 652 tweets. Finally, each county level factor was run against the percentage counterfactual usage and the p-values and R-values were recorded.

**Results**

*Table 1: Counterfactual Usage and Socioeconomic Factor Correlation Experiments at the United States County Level*

<table>
<thead>
<tr>
<th>County Level Factor</th>
<th>R-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Life Satisfaction</td>
<td>0.0471402109547</td>
<td>0.0300912820407</td>
</tr>
<tr>
<td>Divorced Rate</td>
<td>0.0557213788089</td>
<td>0.0117072256707</td>
</tr>
<tr>
<td>Median Household Income (2012)*</td>
<td>-0.0126395369115</td>
<td>0.567732451168</td>
</tr>
<tr>
<td>Percentage Female*</td>
<td>-0.023738526499</td>
<td>0.283157423481</td>
</tr>
<tr>
<td>Median Age (2010)</td>
<td>0.0592168309315</td>
<td>0.00737865312364</td>
</tr>
</tbody>
</table>
Table 1: Shows the correlation experiment results for county counterfactual usage and socioeconomic factors. A negative R-value indicates that as counterfactual usage increases, the factor decreases. *Insignificant values.

Based on the results of the experiments, it can be concluded that many of the correlations are significant based on their p-values. However, it is clear that these correlations are weak with the highest R-value resulting from the violent crime rate experiment. Life satisfaction, divorced rate, median age, poor mental health and Evangelic Protestant Adherent Rates all had positive correlations with increased counterfactual usage. Unemployment, percent under poverty, bachelor's degree rates, Catholic adherent rate, premature death rate, excessive drinking, violent crime, post-secondary education and teen birth rates had negative correlations.

Conclusions and Next Steps

Overall, our deliverables from this project were the model for identifying counterfactuals on Twitter and the framework for obtaining these counterfactuals from the WWBP database. We also created a web-app for demoing this functionality and providing a general overview of counterfactual usage on Twitter. Unfortunately, the results that we obtained from the
county-level statistical correlations were inconclusive, but in the process of studying
counterfactuals we came up with several other applications for our model and framework.
Because counterfactuals are generally associated with planning and decision-making, we could
see our Twitter model being used by politicians to gain insight into what their supporters think
they could do better in their campaign or while they are in office. By identifying counterfactual
tweets having to do with a policy decision, politicians could hear exactly how their supporters
would react to a change in this policy.

On a more granular level, because counterfactual thinking has been shown to relate to
some important aspects of one’s personality, we thought it would be useful for psychiatrists or
doctors to be able to analyze their patients’ social media accounts (with their consent) using our
model in order to ascertain more insights about what type of person their patient is and
therefore be able to help them more thoroughly. Overall, our model has contributed to the
formation of a more concrete definition of what a counterfactual statement is and has expanded
the study of counterfactual thinking into the language of social media for use by the World Well
Being Project.
Works Cited


http://www.cs.cmu.edu/~ark/TweetNLP/#pos.

https://web.cn.edu/kwheeler/grammar_subjunctive.html.