Hashtag Politics:
A Twitter Sentiment Analysis of the 2015 Canadian Federal Election

Amanda Mullins and Adam Epp

Abstract

This paper presents a split plot design model for analysis of sentiment toward federal political parties on the social media platform Twitter in the weeks prior to the 2015 Canadian Federal Election. Data was collected from Twitter's Application Programming Interface (API) via statistical program R. We scored the sentiment of each Twitter message referring to the parties and tested using ANOVA. Our results suggested that the Liberal Party and New Democratic Party had more positive sentiment than the Conservative Party. Actual seat wins coincide with our results for the Liberal Party (which won 148 new seats) and the Conservative Party (which lost 60 seats), but positive sentiment for the New Democratic Party did not correspond to seat wins.

Keywords: Split Plot Design, Confidence Interval, ANOVA, Sentiment analysis

Introduction

Twitter is a popular free-use social media platform where users communicate via short messages of maximum 140 characters called "tweets". In this paper, we propose an approach based on statistical analysis to determine the sentiment with which people talked about federal parties on the social media platform Twitter in the weeks prior to the 2015 Canadian Federal Election.

We consider that if one party appears more than others in tweets with positive words, Twitter users may be expressing positive thoughts and feelings towards that party. This type of data mining (called a "sentiment analysis") is becoming common as social media develops into a major method of

communication in the world. Sentiment on social media platforms like Twitter towards entities like political parties, brands, or companies can be indicative of sentiment of those entities outside of social media as well.

The goal of this study is to find a pattern in the sentiment that Twitter users communicated about the major political parties and leaders in Canada. We designed a statistical experiment such that we can infer that if sampled users of this social media platform speak more positively or negatively about a party, then the population of Twitter users as a whole felt more positively or negatively about that party.

We developed a split-plot model (Montgomery, 2013) for analysis of 140-character messages ("tweets") about the 2015 Canadian Federal Election on Twitter. Our factor of interest is sentiment regarding popular hashtags (a word or phrase proceeded by a hash (#) used to identify messages on specific topics). The hashtags are the subplot factor, while the week of the experiment (performed over three weeks) is the whole plot factor. The experiment is replicated 12 times. Data was collected from Twitter's Application Programming Interface (API) using statistical program R. Using a word lexicon (Hu and Liu, 2004) that attributes positive or negative scores to words, we sum the sentiment of each tweet, and test sentiment of tweets containing hashtags of interest using an Analysis of Variance (ANOVA) test (Montgomery, 2013). In their previous work, Hu and Liu determined their system has good accuracy for predicting sentence sentiment compared to manual classification, with average accuracy at 84% (2004).

The phrase "sentiment analysis" referring to the analysis of evaluative text and using it for prediction was used as early as 2001 (Das and Chen, 2001), primarily with movie (Pang et al., 2002) and product reviews (Turney, 2002). Bermingham and Smeaton (2010) determined that it is easier to identify sentiment in short messages, such as tweets on Twitter, than in longer documents like blogs.

Since the advent of sentiment analysis, there has been

growing interest in using social media to predict the outcome of elections. In a large study of 2009 German federal election, Tumasjan et al. determined that the number of times a party is mentioned in social media is directly proportional to the probability that party will win seats. O'Connor et al. (2010) studied Twitter sentiment in the United States of America and found that sentiment scores correlate with opinion polls on presidential job approval, but that correlation is not as strong for the outcome of elections. The predictive capabilities of sentiment analysis are still debated in literature. Gayo-Avello (2012) suggests that until methods and accuracy for sentiment analysis can be improved, the predictive capabilities are not high. However, sentiment analysis can still be a useful tool for characterizing how social media users feel about political parties.

Statistical Design

We chose to use a split plot design for this experiment (Montgomery, 2013). Split plot designs are common in agricultural studies where, due to logistics, it is difficult to randomize one of the factors. For example, consider a study looking at two types of irrigation on two different types of crops. In a fully randomized design, irrigation and crops should be randomly placed across the fields. It is not feasible to randomize irrigation within a field of crops, due to cost and labour involved. It is, however, relatively easy to randomize the crops. In a split plot design, each field could be split into two plots, one for each type of irrigation, and then the crops could be randomly planted within these subdivisions. The benefit of the split plot design is that there are two levels of experimental units: the whole plot level (which is the "hard to change" factor), and the sub plot level contained within each whole plot. This design only requires

randomization within each whole plot, and not between the

$$y_{ijk} = \mu + \tau_i + \beta_j + \gamma_k + (\tau\gamma)_{ik} + (\beta\gamma)_{jk} + \varepsilon_{ijk} \begin{cases} i = 1, 2, \dots, 12 \\ j = 1, 2, 3, 4 \\ k = 1, 2, 3 \end{cases}$$

τ_i= Replicates

β_j= Tag effect

γ_k= Week effect

$(\tau\gamma)_{ik}$= Whole plot error

$(\beta\gamma)_{jk}$= Tag and Week interaction

ε_{ijk}= Subplot error

whole plots.

Figure 1: Statistical Model

Our statistical model (Figure 1) reflects this design. The response variable y (sentiment score) is modeled as the sum of the overall mean effect, replicate effect, hashtag effect (subplot treatments), week effect (whole plot treatments), whole plot error, effect of interaction between tag and week, and subplot error. The factors and levels are explained below. The proposed model includes interaction between week of study on the hashtag effect and replicate effect. The analysis of the residuals shows that this model offers a good fit to the data, so we chose this more parsimonious model rather than a model with higher level interactions.

This study ran for three weeks prior to the Canadian Federal election. Sentiment might change as Election Day gets closer, so we used week of study as our whole plot factor. Since the split plot design only requires randomization within the whole plots, not between the whole plots, we did not require randomization across the weeks.

Data collection for this study was performed using the Twitter API, which allows read access to tweets. To obtain the data we used the TwitteR R package (Gentry, 2015) which

provides an interface to the Twitter API. Using the TwitteR package, we were able to pull tweets based on the words contained in the tweets. We used the hashtags #NDP or #ThomasMuclair, #LPC or #JustinTrudeau, #CPC or #pmharper, and #cdnpoli. We determined that these were the most popular hashtags at the time of the election ("Canadian Politics Twitter Hash Tags", n.d.). Each time the program ran, it collected 50 tweets for each hashtag. Those tweets were processed through an algorithm that checks the number of positive words minus the number of negative words in the tweet. There are two text files developed by Hu and Liu (2004): one that contains positive words and one that contains negative words, which we used to apply a numerical score ("Sentiment Score") for each tweet. We then summed up all 50 of the tweets to get a number that represents the sentiment for a political party for a specific time. Our response variable is the Sentiment Score and is calculated as the average score for these 50 tweets at that given level.

The treatments we are most interested are for our subplot factor hashtag. The program ran, collected the 50 most recent tweets using the hashtag of interest, and then stopped. The time between hashtag searches was a matter of seconds. Twitter users send out tweets at whim, and thus the Twitter users themselves effectively randomized the sampling of tweets with the hashtags of interest. This information is summarized in Table 1.

Table 1: Factors and Levels

Factor	Level
Hashtag (subplot factor)	#cdnpoli (Level 1) #cpc or #pmharper (Level 2) #ndp or #thomasmulcair (Level 3) #lpc or #justintrudeau (Level 4)
Week of Experiment	Week 1 (Level 1)

Table 1: Factors and Levels	
(whole plot factor)	Week 2 (Level 2) Week 3 (Level 3)

Both treatment effects for the whole plot and for the subplot are fixed, and only replicates are random (hence, so are interactions involving replicates). The sentiment score could be influenced by the location of tweeters (for example, due to its conservative voting history, people in Alberta may be more likely to favour conservative parties compared to people in Quebec). To account for this, we pulled sums from different times (on the hour from 1:00 to 4:00 PM) to obtain a better sampling of people across the country as time zones differ.

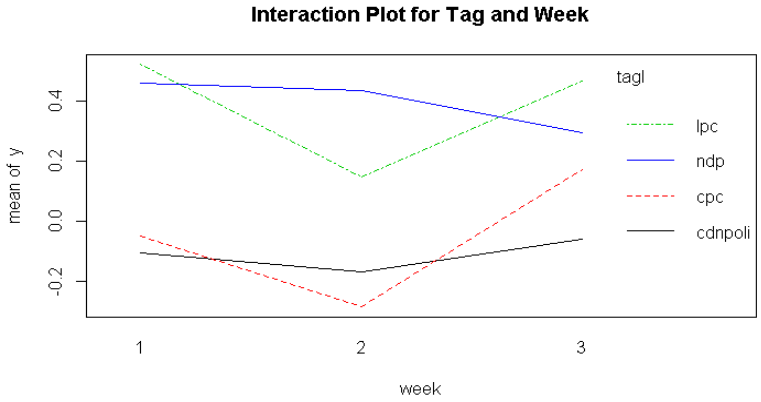
Days of the week are also likely to influence the sentiment score (people may be more or less likely to respond positively during different times of the week). In order to minimize this effect, we ran the program on three different days each week (Tuesday, Thursday, and Sunday). Thus, the experiment was replicated 12 times over the course of the study. We were limited due to time constraints on how many replicates were feasible. In total, we collected 7,200 tweets (600 per hashtag, per week.)

Statistical Analysis

We conducted an analysis of variance for the collected data. To determine whether our model and design were appropriate for this data, we analyzed interaction plots and did a residual analysis.

Figures 2 and 3 show the interaction plots for tag and week, and replicates and week, respectively. From these plots, we determined that there is interaction between week of study on the hashtag effect and the replicate effect. This is in agreement with the proposed split-plot model.

Figure 2: Interaction Plot for Tag and Week



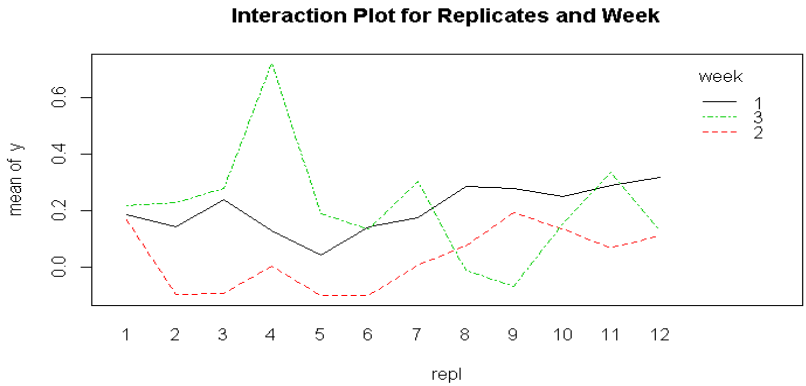


Figure 3: Interaction Plot for Replicates and Week

The ANOVA table in Table 2 shows that the effect of hashtag

on sentiment is significant, with a p-value of 3.38×10^{-16}. It also shows that there is significant interaction between week and tag, with a p-value of 0.004644. However, since the ANOVA calculations in R assume that all factors are fixed factors, and our replicates are a random factor, the F-statistic calculated in the ANOVA is not the appropriate statistic. Instead of using the mean square of the residuals, we calculate it using the mean square of the interaction of replicates and week (Montgomery, 2012). This gives an F-statistic of 6.040427, which corresponds to a p-value of 0.00811025. Therefore, we can conclude that the effect of the week on the sentiment is also significant.

Table 2. Analysis of Variance

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
repl	11	0.6316	0.05741	0.7984	0.641207	
week	2	1.0519	0.52594	7.3141	0.001090	**
tag	3	8.0339	2.67798	37.2419	3.38e-16	***
repl:week	22	1.9155	0.08707	1.2108	0.256792	
week:tag	6	1.4512	0.24187	3.3636	0.004644	**
Residuals	99	7.1188	0.07191			

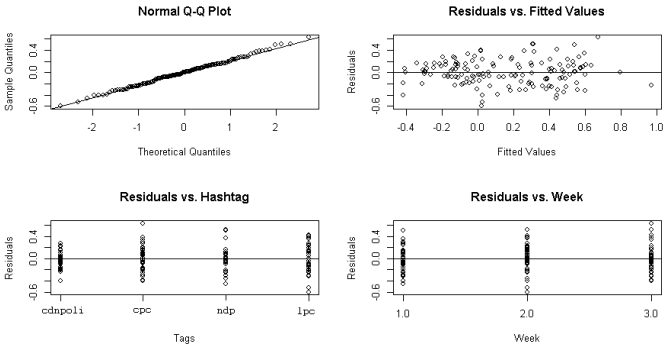
Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
> MSweek <- 0.52594
> MSRepWeek <- 0.08707
> Fweek <- MSweek/MSRepWeek
> Fweek
[1] 6.040427
> pweek <- (1-pf(Fweek,2,22))
> pweek
[1] 0.00811025
```

Figure 4 shows the residual analysis of the data. The response variable follows an approximately normal distribution, as per the relative linearity in the Normal Q-Q plot. The Residuals vs. Fitted Values plot indicates the residuals and fitted values do not dramatically deviate from homoscedasticity. When plotted against the subplot treatments for hashtags and whole plot treatments for weeks, the residuals show the same relatively even spread on the y-axis, indicating approximately constant variance. Bartlett and Levene statistical tests also showed homoscedasticity of residuals. Thus, the data does not have any major violations of the design assumptions.

Figure 4: Residuals Analysis

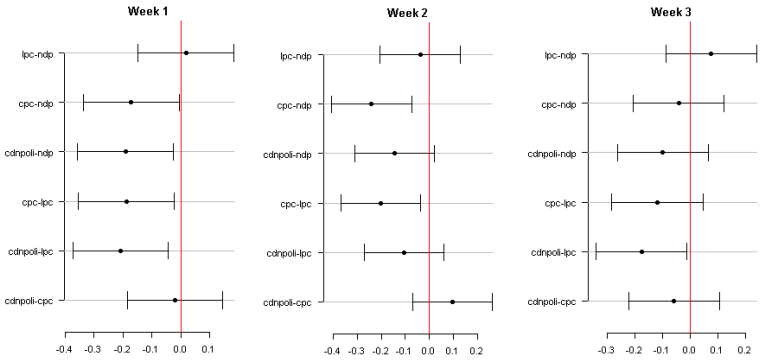


Multiple Comparisons

After the ANOVA tests we conclude that the effect of the hashtag treatments on sentiment is significant, which indicates that Twitter users use more positive words when speaking about some parties than others. However, since the ANOVA also shows that the interaction effect of weeks and hashtags is significant (and thus, the sentiment to which people spoke of parties varied from week to week), to determine which parties have more positive sentiment, they must be compared on a week-by-week basis.

We produced a 95% Tukey-Kramer multiple comparisons confidence intervals for the hashtag treatments at each week level. The results are plotted in Figure 5. The 95% confidence intervals for the difference between each hashtag treatments are plotted horizontally. If the interval does not cross the red line at zero, we are 95% confident that the difference between the hashtags treatments is significant.

Figure 5: 95% Tukey-Kramer Multiple Comparisons



For Week 1 both the #NDP hashtags and #LPC hashtags scored significantly higher than the #CPC and #cdnpoli hashtags. Thus, in average, Twitter users used more positive words in tweets with #NDP and #LPC hashtags, compared to those with #CPC and #cdnpoli hashtags. There is no difference in sentiment between the #NDP and #LPC hashtags, nor between the #CPC and #cdnpoli hashtags.

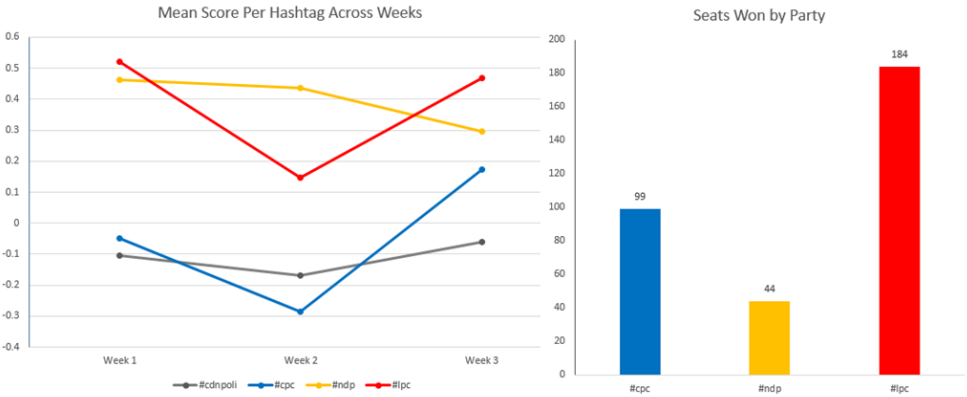
The plot for Week 2 shows that both #NDP and #LPC hashtags on average scored higher in sentiment than #CPC. There was no difference in sentiment for #NDP and #cdnpoli, or for #LPC and #cdnpoli. There was also no difference in sentiment for #CPC and #cdnpoli.

For Week 3 in average #LPC scored higher in sentiment than #cdnpoli. However, there was no difference in sentiment for any of the other hashtag treatments. Interestingly, Week 3 was the week of the election, and it showed no difference in sentiment between any of #NDP, #LPC, or #CPC.

Conclusion

The 2015 Canadian federal election resulted in a change of government with the Liberal Party gaining 148 seats, the Conservative Party losing 60 seats, and the New Democratic Party losing 51 seats. We hoped that a Twitter sentiment analysis conducted in the weeks prior to the election would illustrate how people on Twitter spoke about the political parties and mirror the seat wins. Our analysis partially reflected the results of the election, as seen in Figure 6.

Figure 6: Hashtag Sentiment and Seat Wins



Throughout the weeks, in average the hashtags used for Canadian politics in general (#cdnpoli) and for the Conservative Party had lower sentiments than the other hashtags. Twitter users were using more negative words in tweets with these hashtags than tweets with other hashtags. This implies that users were experiencing negative feelings in tweets about Canadian politics and the Conservative Party. It is understandable that these hashtags were similar in sentiment across the weeks, as when people spoke of Canadian politics in general, they may have been referring to the current party in control of the government before the election – the Conservative Party. These negative feelings may have contributed to the loss of 60 seats that the Conservative Party experience.

In contrast, the Liberal Party hashtags and the New Democratic Party hashtags had higher sentiments, indicating that Twitter users were using more positive words in tweets containing these hashtags. The positive feelings towards the Liberal Party may explain why the Liberal Party had a huge comeback, winning 148 seats for a majority government. However, the positive sentiment on Twitter did not correlate to seat wins for the NDP.

There is a bias in just sampling Twitter users. Social media users tend to be younger, and younger people tend towards more liberal views. This type of sentiment analysis may benefit from including factors such as age or location of users in future work. Due to schedule constraints, we were not able to sample at times aside from the afternoon. There may have been bias from sampling just afternoon Twitter users, and further analysis of this type may benefit from randomization of time of sampling.

Other opportunities for sentiment analysis could include analysis of future elections across the world. From our experiment, we anticipate interesting results from data mining of social media with larger samples, more replicates, and over longer periods. Applications of sentiment analysis can include social media presence monitoring, which is useful for political parties, corporations, or even individuals. If a solid model can be developed, social media sentiment analysis may even be useful for predictions.

Our model includes an interaction of the week of study and the hashtag factors. The sentiment associated with each hashtag treatment changed from week to week. This makes sense, even in Canada's longest election campaign since 1872, at 78 days. As Election Day gets closer, parties will make their last push at campaigning, citizens become more engaged, and media will publish new scandals and information. Our model provides an interesting snapshot of sentiment of Twitter users regarding Canadian political parties in the weeks prior to the election.

Acknowledgements

Preliminary versions of this paper were presented at the 2016 MacEwan Research Week, the 2016 Undergraduate Research in Science Conference of Alberta (URSCA), and the 2016 Statistical Society of Canada Student Conference.

This work was partially supported by a grant from MacEwan's Undergraduate Student Research Initiative (USRI). We would like to thank Prof. Cristina Anton for mentorship for this project.

References

Bermingham, A., & Smeaton, A. F. (2010). Classifying sentiment in microblogs. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management - CIKM '10*.

Canadian Politics Twitter Hash Tags. (n.d.). http://politwitter.ca/page/canadian-politics-hash-tags

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science, 53*(9), 1375-1388. http://doi.org/10.1287/mnsc.1070.0704

Gayo-Avello, D. (2012). "I wanted to predict elections with Twitter and all I got was this lousy paper". *CoRR*. abs/1204.6441.

Gentry, J. (2015, July 29). TwitteR package - The Comprehensive R Archive Network. https://cran.r-project.org/web/packages/twitteR/twitteR.pdf

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '04*.

Montgomery, D. C. (2012). *Design and Analysis of Experiments* (8th ed.). New York: Wiley.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *Proceedings of the International Conference on Weblogs and Social Media (ICWSM).*

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - EMNLP '02*.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape. *Social*

Science Computer Review, 29(4), 402-418. http://doi.org/10.1177/0894439310386557

Turney, P. D. (2001). Thumbs up or thumbs down? *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*.