Computational Sociolinguistics of Scottish Twitter and the Independence Referendum

Debnil Sur
Department of Computer Science, Stanford University
debnil@stanford.edu

Abstract

Understanding the relationship between sociolinguistics and politics on Twitter can help us understand how social media activity correlates with real-world political events. We focus on Twitter activity about the 2014 Scottish Independence Referendum. A major international political event, this was the subject of much social media activity and thus provides a fascinating case study of this relationship. We curate the first known dataset of Scottishisms using Twitter; study whether Twitter users modulate their use of Scottish slang when discussing politics; and quantify the relative use of slang between supporters and opponents of the referendum. This case study helps us better understand Scottish Twitter's sociolinguistics and its variation between major political factions. Through this case study, we hope to move towards a better understanding of the relationship of sociolinguistics and politics on Twitter and eventually to how those forces affect political change.

1 Introduction

Social media are increasingly being used to discuss contentious domestic political issues. While the most well-known examples come from protests, such as the Arab Spring and the #BlackLivesMatter movement, elections, televised debates, and other more conventional political events are also the subjects of significant discussions on these platforms (Flores, 2015), (Wang et al., 2012). A parallel trend is one in language: social media users use both Standard English and their own regional colloquialisms, in addition to creating new linguistic forms of their own (Tagliamonte and Denis, 2008). The micro-blogging platform Twitter exhibits both of these trends: hashtags are often used to discuss political issues (Small, 2011), and demographic groups heavily use existing slang (Kouloumpis et al., 2011). It thus provides an excellent forum to study the intersection of these trends in political discussion and regional colloquialisms. True to the theme of this seminar, we focus on a regionally relevant political event which has gained importance in the wake of recent domestic political events: the 2014 Scottish independence referendum. We will study the use of "Scottishisms" on Twitter in general and as it pertains to this event. By "Scottishisms", we denote words from Scots, Scottish variants of Standard English, and particular slang that has risen on Scottish Twitter.

Following the above motivation, we ask three major research questions.

RQ1. Is there a significant use of Scottishisms on Twitter in Scotland?

RQ2. Do people tweeting about the referendum use different frequencies of Scottishisms than in their normal Twitter activity?

RQ3. Does one side of the Scottish independence use more Scottishisms?

We hope to utilize this work as a case study to provide examples answering the following questions, in turn. First, the occurrence of regional dialect on Twitter. Second, the relationship between political topics and general language, from a neutral perspective, can show how users modulate their voice based on topic and intended audience. Third, the relationship between ideological difference and use of slang helps extend existing theories of political language use into the Twittersphere.

2 Related Work

Each of the three research questions has been briefly covered in the existing literature. Before delving into specifics, we will note that no prior computational work has utilized Scottish Twitter data. While this made finding relevant work and developing an intuition for our results more difficult, it does guarantee some novelty to our eventual findings.

We began this paper by noting the use of social media in political protest and organization. Work to this end has been almost wholly conceptual, and computational approaches have users with political intent rely on the hashtag for two purposes: organizational (to facilitate description or navigation) and social (to facilitate resource exposure and sharing) (Yang et al., 2012). Two aspects of the hashtag separate it from traditional social tags. First, it has a generative role not seen in conventional social markers. For instance, scientific conference communities in Twitter have been argued as built on the adoption of one specific hashtag (Ebner and Reinhardt, 2009). A hashtag therefore identifies and forms a community and lets users join it - traditional social tags do not play such a defining role. Second, the conversational nature of Twitter makes its hashtags much richer than their equivalents on other social media platforms (Yang et al., 2012). Twitter's conversational nature makes its user interactions far more frequent, and hashtagging posts therefore creates more detailed inner-community conversations. Tracking a certain hashtag thus provides a richer profile of actual conversations between users of a certain political bent. Its data is thus ripe for studying differences within these groups.

The use of regional dialects on social media has also been documented significantly. Past work has found that lexical variation in tweets is used to encode phonetic differences in user language (Eisenstein, 2015). Particularly with respect to African-American Twitter, users have been found to use regional slang in both regular and conversational tweets. These differences have been sufficiently pronounced that unsupervised probabilistic models have latently determined regional dialectic differences among African-American Twitter users from different major American metropolitan areas (Hong et al., 2012). The only work with respect to Scottish Twitter data has used a few dozen tweets from specific users and shown the existence of the 'ae' for 'o' substitution that most strongly characterizes the Scottish accent (i.e., 'cannae' for 'cannot', 'dae' for 'do', or 'tae' for 'to') (Tatman, 2015). There have been no large-scale, computational attempts to characterize the Scottish dialect on Twitter.

Similarly, almost no computational linguistic approaches have been taken with respect to the Scottish Independence referendum. Held on 18 September 2014, the ballot initiative determined whether Scotland would leave the United Kingdom and form its own independent nation. With a turnout of 84.6%, it had wide-reaching implications and was accordingly widely discussed. The only other computational approach to studying the referendum utilized Twitter data to develop a topic-based Naive Bayes text-classification schema for political orientation of individual tweets (Fang et al., 2015). While computationally interesting, this answers a very different social question than ours.

Finally, while the significance of regionalisms in political discussion has rarely been studied using social media or computational techniques, it is well-documented in sociolinguistics. The use of dialects commonly connotes appeals to an in-group or common identity, while using a more standard speak can indicate a desire to appear rational or well-educated (Sauter, 2013). Though this interpretation has not been rigorously validated using dialect patterns on social media, it is one possible explanation to keep in mind when analyzing our eventual results.

3 Method

3.1 Data Collection

3.1.1 Tweet Collection

We started with a set of approximately 170 million tweets from September 2013 through September 2014. The date range was chosen to cover a year's worth of activity before IndyRef (on September 18, 2014). The data was collected using the Twitter Streaming API, which collects a subset of Twitter activity in near-real time. Unlike the full Firehose API, this API provides a random subset of 1% of all activity. The research group had already discarded non-English tweets from this subset, giving us the set of 170 million tweets we began with. From this set, we immediately filtered out retweets, as their presence gives no useful sociolinguistic information about the actual tweeter.

Hashtag	Side	Frequency
IndyRef	Neutral	60,357
VoteYes	Yes	13,864
YesScotland	Yes	1,820
VoteNo	No	2,760
BetterTogether	No	4,961

Table 1: The hashtags used in the Referendum campaign, with orientation and frequency in our dataset.

They simply provide information identical to that of other users, whether in the dataset or not, and are therefore not useful for studying original use of dialect by the actual tweeter. This filter returned approximately 120 million tweets. Moving forward, we refer to this dataset as the 'total dataset'.

We then had two major filtering tasks to generate our desired datasets. The first was finding relevant geotagged data. We collected 16,439,214 geotagged tweets by finding tweets with non-zero latitude and longitude fields in our total dataset. We then filtered this geotagged data to approximately 170,000 from Scotland and 1,824,782 from the United Kingdom as a whole. The second was finding tweets about the referendum. We did this using the five major hashtags used in the referendum campaign: #IndyRef, which was politically neutral; #VoteYes and #YesScotland, the 'Yes' hashtags that favored independence; and #VoteNo and #BetterTogether, the 'No' hashtags that favored staying in the United Kingdom. Their frequencies are shown in Table. 1.

3.1.2 Finding Scottishisms

Using the geotagged tweets, we find words more frequently used in Scotland than in the world at large. A major challenge in this search is the absolute sparsity of Scottish-specific words. A quick manual inspection of the geotagged data shows that even the most well-known examples of Scottish slang are used absolutely rarely and relatively far less frequently than their Standard English counterparts (Table 2). Finding statistical significance from such sparse counts is difficult, and the output of basic probabilistic models of frequency were dominated by non-sparse words that occurred more in Scottish tweets (such as Edinburgh, Rangers, or Glasgow) rather than the far sparser colloquialisms.

To correct for this sparsity, we run the Sparse Additive Generative Model (Eisenstein et al., 2011). We chose this model because it handles sparsity well, and its original case study involved finding differences in regional dialect in a geotagged corpus. The SAGE model is meant to correct a problem with the overreliance of generative models of text on the Dirichlet-multinomial conjugate pair, in which the Dirichlet prior contributes pseudocounts to the observed counts generated by the multinomial. The ease of parameter estimation results in complicated latent variable structures and a lack of robustness when training data is limited. Eisenstein et al. identify these problems as stemming from directly modeling the lexical probabilities associated with each document class. On the other hand, SAGE models the difference in log-frequency from a background lexical distribution. This lets us limit the number of terms under consideration and increases model scalability by taking advantage of the additivity of logarithms.

Scottishism	Count	English	Count
aboot	108	about	5853
cannae	404	can't	3521
dae	125	do	5878
gawn	46	go	3898
nae	292	no	6346
wi	189	with	11019
wit	226	what	5275
yi	58	you	2331

Table 2: Classes, examples, and frequencies of words disproportionately occurring in Scottish tweets.

As input for this model, we used a background probability distribution which was all words that occur more than fifty times in the set of all geotagged tweets, with stopwords, hashtags, URLs, and @-mentions removed. For the counts vector, we input the number of times each of those words occurs in the set of geotagged tweets from Scotland. The model then shows us which words are used relatively more frequently in Scotland when compared to the dataset as a whole. When manually perusing the returned data, we realized that many words that frequently occurred were British words or proper nouns (such as 'lorry' or 'London') and therefore did not directly identify Scottishisms. We then modified our inputs to be a background vocabulary with similar restrictions from British geotagged data and the corresponding

Class	Examples	Frequency
Proper Nouns	Edinburgh, Salmond, TITP	147
Standard English	reprobate, patched, rainfall	84
Syntactic	yerself (yourself), doesny (doesn't), oan (on)	47
Semantic	burd (bird), hame (home), boke (vomit)	91
Slang/Misc.	ooft, belters, jst	42

Table 3: Classes, examples, and frequencies of words disproportionately occurring in Scottish tweets.

counts of these words in the Scottish geotagged data.

Using this second output, we find a set of 411 words used significantly more in Scotland than in the UK as a whole. We stopped including words in our dataset at a log-deviation of 1.5; at this point, many words included were English words disproportionately used in Scotland, rather than Scottish words. We classified these high-performing words into five groups: proper nouns, standard English words, Scottish syntactic words, Scottish semantic words, and miscellaneous words/slang. Examples of each and their relative counts are shown in Table 3.

. Given that we wish to detect differences in the use of regional slang, we chose to pursue common Scottish words that had a clear variant in Standard English. The use of Scottish proper nouns, like Edinburgh or SNP, does not help us identify when speakers actively choose to use a Scottish term over its standard English form. For similar reasons, we do not consider miscellaneous words or Scottish-specific slang, as they have no clear English variant, and also discard English words that appear disproportionately in Scotland. Of the remaining, we omitted words that are used less than 50 times in the set of all Scottish tweets. This left us with 85 of the original 411 Scottish words. As a baseline check, we found a ratio of 0.06 over all geotagged tweets with the chosen words; that is, for every 6 occurrences of Scottish words, its English variants occurred 100 times.

3.2 Experiments

3.2.1 User Modulation

The prior steps generated five sets of hashtag-based tweets. We want to find out if users tweeting about the referendum use more or fewer Scottishisms when discussing this event, as opposed to their general usage. Though we collected a variety of hashtags, we choose to focus on the neutral political hashtag #IndyRef; this avoids the potential confound of ideological bias, as voters with a strong political leaning on either side may differ from more neutral observers in the frequency and type of Scottishisms used. To generate a control set, we collect all tweets from the same users not using any of the referendum hashtags. If a user in the original data set has no corresponding tweets in the control set, we remove their tweets from the #IndyRef set, as direct comparison of dialect use is impossible. This set of 1,110,335 tweets also overrepresents users who most frequently tweet. To correct for this, we downsize this large control set to construct a one to one smaller control set: for every tweet by a certain user containing #IndyRef, we include a tweet by the same user without any referendum hashtags. This leaves a smaller set of 55,219 control tweets, matched one-to-one by author with a set of #IndyRef tweets. If our dataset contains 5 tweets with the #IndyRef hashtag by a user and 50 tweets by that user without any referendum hashtags, the final balanced dataset would contain those 5 #IndyRef tweets and 5 of her non-referendum tweets as well.

Recall that we wish to compute the difference in use of Scottish common words between #IndyRef tweets and the control set. To do this, we calculate the aforementioned summary statistic for each distribution: $\frac{\text{\# times Scottish words used}}{\text{\# times Standard English forms used}}$. We then take the difference in these values between the two sets and use this as our target variable. This gives us the true statistic. For interpretation, note that 0 difference would mean that Scottish Twitter users use the same amount of regional dialect in their general tweets and tweets about the referendum.

3.2.2 Monte Carlo Permutation Test

Once we calculate this difference, we wish to determine its statistical significance. We conduct a Monte Carlo permutation test. In general, to conduct a permutation (or randomized) test between two lists L_1 and L_2 (of size n_1 and n_2), we concatenate the lists; shuffle their values; and split the shuffled, larger list into smaller lists of sizes n_1 and n_2. We then recalculate the difference between the ratios in these modified lists. Note that ideally, we would consider every possible permutation; however, due to the size of the IndyRef list (55,219 tweets), there are over 10^{8000} such rearrangements. Computational limits, time, and feasibility make sampling all of these infeasible. Instead, we take a Monte Carlo approach and use 100,000 permutations to generate a distribution of differences in the Scottish/English ratio between the two sets. We calculate the p-value by finding the proportion of simulated differences with absolute value greater than the true statistic. The percentage of values more extreme than the true difference gives the probability that this could occur randomly.

We use a similar approach to testing the difference in the frequency of use of Scottish dialect between supporters and opponents of the referendum. We construct two large datasets. The first contains tweets supporting a 'Yes' vote (#VoteYes, #YesScotland) and the second tweets supporting a 'No' vote (#VoteNo, #BetterTogether). Unlike our previous controlled sample, these lists are dissimilarly sized, with about 19,000 tweets in the first and 9,000 in the second. Nonetheless, our procedure can work with input lists of different sizes, so we conduct a permutation test using 100,000 measurements to test significance.

4 Results and Discussions

Let us now review the results of our data collection and experiments with respect to each of the research questions posed at the start of our paper.

4.1 Scottishisms on Twitter

First, are there Scottishisms on Twitter? Yes; we found significant evidence of Scottishisms used on Twitter, broadly divided into five categories: proper nouns, English words, syntactic words, semantic words, and miscellaneous words and slang. It is of note that the absolute percentage of tokens in the geotagged Scottish data that are Scottish words is 14.5%. When considered in comparison to English counterparts, as our experiments do, this ratio only becomes smaller. Thus, while the Scottish regional dialect is well-known and well-documented, its absolute use on Twitter is quite sparse.

However, their use in Scotland, when compared to a distribution of all words used in the United Kingdom, was regarded as statistically significant by the SAGE model. This model yielded 411 words. Of these, we identified 85 words as our dataset for the following experiments. These were Scottish semantic and syntactic words that occurred in non-sparse quantities and had clear English variants. While we discarded many of the words we found due to our specific experimental needs, this does provide a useful set of words for future studies of Scottish Twitter.

4.2 Regionalisms and Politics

Second, do Scottish users modulate the frequency of Scottishisms in their tweets about the Referendum? Yes; we found strong evidence that users slightly modulate their voice based on the population and topic. In an absolute sense, this difference is very small. We found a Scottishism ratio of 0.0165 in general conversation and of 0.013 when discussing the referendum. Two related findings yield this difference its significance. First, the p-value yielded by the Monte Carlo permutation test was extremely small ($p < 0.001$). This demonstrates a very low probability that this difference could occur by chance and thus lends it statistical significance.

Second, the baseline statistic was absolutely small: these words had a proportion of 0.06 with respect to their variants in the set of all geotagged Scottish tweets. Thus, their use was not especially high in a set of tweets explicitly from Scotland, and being lower than this benchmark is not surprising for a political event with major international implications. Moreover, the presence of foreign Twitter users also discussing #IndyRef could have diluted this statistic. As an extension of this idea, not all users using #IndyRef use any Scottishisms. We considered taking such users out of the sample, because if they do not use such phrases at all, studying how they modulate their voice is impossible. Unfortunately, doing so would have left us with too few remaining users and would let the data drive the experiment design too much.

Leaving these users in still yielded a statistically significant result, thus justifying their presence.

The conceptual implications of this finding are slightly unclear. Though statistically significant, the small magnitude of both their absolute use and their difference makes meaningful generalization difficult to support. Moreover, unlike most prior work, we based our study around a political marker explicitly designated as neutral in the #IndyRef hashtag. While there are clear interpretations for the use of regional rhetoric in political contexts among people of certain ideological bents, the reason behind the decreased deployment of such terms in a neutral setting has not been studied in prior work on political language.

4.3 Regionalisms and the Referendum

Third, does one side of the referendum debate use more Scottishisms? Yes; we found similarly strong evidence that pro-independence voters use Scottish words in their tweets about the referendum statistically significantly more than pro-Britain users. Again, the absolute values of these uses are quite small: 0.017 for pro-independence voters and 0.009 for pro-Britain users. However, as before, the p-value generated by the Monte Carlo permutation test is extremely small ($p < 0.001$), demonstrating the significance of this difference.

Moreover, the small baseline ratio of 0.06 further accentuates this difference, as it provides a likely upper bound on our statistic. As before, we would be highly surprised if there was a greater occurrence of Scottishisms in a political event with global implications than in explicitly Scottish data. Given this baseline, it becomes far more telling that those who supported Scottish independence were twice as likely to use the native language than those opposed, even if the absolute proportions of use are small. This provides further evidence that the use of dialects in politics often indicates appeals to in-group identity and nationalism. It extends this existing theory of language use into the social media sphere.

4.4 Future Work

Future work should proceed in three directions.

Run more simulations and tests. First and foremost, we need to run more simulations and tests to confirm these results. In particular, while $100,000$ runs of the Monte Carlo estimation method let us determine a reasonable estimate, it is still a small fraction of all possible permutations of our data. More time and computational resources will help further validate the method we have already used. It will also be useful to validate this finding with alternate re-sampling methods, like the bootstrap and jackknife, to ensure that our result can be found via similarly motivated methods with different approaches.

Study the granularity of the difference. After identifying a statistically significant high-level finding, we wish to identify the user attributes that most strongly correlate with the use of Scottishisms in political contexts. This can occur through either a supervised or unsupervised approach. For the former, many Twitter users have personal data stored on their accounts, such as age, gender, and region. We can use this information to divide the final datasets and study how the use of Scottishisms differs between these groups.

For the latter, unsupervised algorithms using geotagged data have successfully identified geographical lexical variation in social media. Others have been able to identify age and gender from tweets. A combination of these models could potentially have the same division as the supervised approach. Following both paths would naturally provide a comparative study about the efficacy of supervised and unsupervised approaches in quantifying differences in the use of Scottishisms.

In future work, we will also vary our target variable. Rather than calculating the summary ratio statistic that we did in this paper, we will also study the use of Scottishisms in more absolute terms. This will show how specific political and demographic factions use certain types of Scottishisms (i.e., syntactic or semantic words) compared to others and provide a more accurate picture of the difference in social media use among Scots.

Compare Twitter with votes. To understand the real-world implications of this and other studies of Twitter use, we must match up the demographic results we find with actual voting patterns. How do the Twitter behaviors of various social, economic, or political factions compare with their actual voting? Did regions of Scotland that supported independence on Twitter support it at the ballot box? This and similar questions will help network scientists, campaign advisors, and many other academics and policymakers better understand the relationship between social media and

electoral results. Better answering that question can profoundly shape the future of democracy as we know it.

4.5 Acknowledgments

We would like to thank Prof. Sharon Goldwater and Pippa Shoemark at the University of Edinburgh for their guidance in the data collection, experimental design, and interpretation of results. Without their help, this project would have gone nowhere, and their advice in data collection and management has been particularly instructive in learning how to answer interesting questions with large data sets. We would also like to thank Luke Shrimpton for his help in curating the original Twitter data and dealing with difficulties along the way. Finally, we would like to thank Prof. Mykel Kochenderfer, Rachael Tompa, the Bing Overseas Studies Program staff, and all our fellow Bing Scholars for making this such a wonderful educational experience and research opportunity.

References

Martin Ebner and Wolfgang Reinhardt. 2009. Social networking in scientific conferences–twitter as tool for strengthen a scientific community. In *Proceedings of the 1st International Workshop on Science*, volume 2, pages 1–8.

Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text.

Jacob Eisenstein. 2015. Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, 19(2):161–188.

Anjie Fang, Iadh Ounis, Philip Habel, Craig Macdonald, and Nut Limsopatham. 2015. Topic-centric classification of twitter user's political orientation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 791–794. ACM.

Joseph L Flores. 2015. *# blacklivesmatter: The investigation of twitter as a site of agency in social movements*. Ph.D. thesis, THE UNIVERSITY OF TEXAS AT EL PASO.

Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J Smola, and Kostas Tsioutsiouliklis. 2012. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st international conference on World Wide Web*, pages 769–778. ACM.

Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! *Icwsm*, 11:538–541.

Disa A Sauter. 2013. The role of motivation and cultural dialects in the in-group advantage for emotional vocalizations.

Tamara A Small. 2011. What the hashtag? a content analysis of canadian politics on twitter. *Information, Communication & Society*, 14(6):872–895.

Sali A Tagliamonte and Derek Denis. 2008. Linguistic ruin? lol! instant messaging and teen language. *American speech*, 83(1):3–34.

Rachael Tatman. 2015. # go awn: Sociophonetic variation in variant spellings on twitter. *Working Papers of the Linguistics Circle*, 25(2):97–108.

Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120. Association for Computational Linguistics.

Lei Yang, Tao Sun, Ming Zhang, and Qiaozhu Mei. 2012. We know what@ you# tag: does the dual role affect hashtag adoption? In *Proceedings of the 21st international conference on World Wide Web*, pages 261–270. ACM.