Identifying Opinions of Toronto Neighborhoods from Twitter Using Sentiment Analysis

# Introduction

The way neighbourhoods are perceived by its residents or the general public plays an important role in understanding measures like neighbourhood quality, residential mobility patterns, the price of homes and the general well-being of residents (Permentier & van Ham, 2011). City initiatives like Edmonton’s Neighbourhood Revitalization Program understand that well-built neighbourhoods support residents’ quality-of-life.[[1]](#footnote-1) These initiatives look to engage with citizens to determine which neighbourhoods need to be prioritized for renewal.

One of the challenges that governments and organizations face with neighbourhood revitalization initiatives is in collecting and analyzing the measures or evidence that can be used to prioritize which neighbourhoods are most in need of investment (Strong Neighbourhood Task Force, 2005). A tool that is sometimes used when doing this type of planning is a civic survey. For example, in 2018, the City of Saskatoon put out their Civic Services Survey: Performance, Priorities & Preferences to gather insights on residents’ experience/impressions related to the quality of areas such as community and public services and recreation and culture services.[[2]](#footnote-2)

Though civic surveys are useful ways to obtain residents opinions and data, there are emerging ways to gather the information needed that can be used to prioritize neighbourhoods by utilizing the stores of “big data” available on the internet. Specifically, by analyzing Twitter posts or *tweets*. Twitter contains vast amounts of publicly accessible data that is produced by its approximately 126 million active users as reported in 2019.[[3]](#footnote-3) Justin Hollander, an associate professor of urban and environmental policy and planning at Tufts, running the Urban Attitudes Lab[[4]](#footnote-4) believes that this type of data can be used to inform planning and policy and help us understand our cities (Tuhus-Dubrow, 2014, October 21).

This paper will be exploring the field of Sentiment Analysis (SA), which focuses on analyzing the polarity of a given text, document or phrase categorizing and assigning the content a positive or negative sentiment (Ankit, 2018; Giachanou & Crestani, 2016). SA methods will be applied to tweets pulled from Twitter to identify the polarity of sentiments that people have towards different Toronto neighbourhoods. In presenting a practical approach to weighing and ranking sentiments of these neighbourhoods, this paper hopes to demonstrate the value that Twitter Sentiment Analysis (TSA) and SA of other social media platforms can have in informing revitalization initiatives, and areas such as city planning and policy development.

# Literature Review

Given that Twitter is a social phenomenon with vast global reach producing seemingly endless streams of data, it is no wonder that TSA is a field that has attracted much research interest in data science. Giachanou & Crestani’s (2018) comprehensive survey*, Like It or Not: A Survey of Twitter Sentiment Analysis Methods*, helps paint a broad stroke of the studies that have carried out sentiment analysis on Twitter data, categorizing them according to their approaches. In addition, Giachanou & Crestani (2018) are able to succinctly outline all the challenges that come with developing SA methods specific to Twitter. They also introduce and list important types of feature selection and review the most frequently used evaluation metrics for TSA which will be addressed in this paper (Giachanou & Crestani, 2018).

Ankit’s (2018) paper, *An Ensemble Classification System for Twitter Sentiment Analysis,* proposes a methodology that opposes the use of one best classifier for SA and instead offers a methodology for the use of an ensemble classification system. The system proposed uses base classifiers, Naïve Bayes, Random Forest, Support Vector Machine and Logistic regression to get comparative accuracy, precision, recall and F1 scores (Ankit, 2018). Ankit (2018) then provides the ensemble algorithms which take the predictions of the base classifiers to enhance the performance and accuracy of base learning techniques. Ankit’s (2018) results show that against the base classifiers the ensemble classifier proposed performed better than the stand-alone classifiers. The ensemble classification system Ankit (2018) outlines will be applied in this paper.

Ankit’s (2018) paper used Bag-of-words technique to convert tweets into numeric representation. In Bouazizi et.al.’s (2016) paper, the authors use Term Frequency-Inverse Document Frequency (TF-IDF) since they have a large text to mine and many features would be produced to select the top features to use in by weighting and scoring the frequency of words in the text. This technique will be used in this report as we are dealing with a large volume of tweets and features.

As popular a research topic TSA is in data science, few studies were found that utilize it to obtain insights into the health, development and vitality of city neighbourhoods and its residents in the social sciences. Studies like Saeidi et. al.’s (2016) and Hu et. al.’s (2019) utilize neighborhood reviews from review specific platforms and apply SA to their datasets. Gibbons et. al. (2019) and Nguyen et. al. (2017) are the few papers that look at TSA at the neighbourhood level to predict and examine resident’s health. This paper intends to bring light to the value TSA can bring when applied in social science contexts.

# Dataset

Twitter tweets are used as a data source in this paper. The data was pulled from Twitter using Tweepy[[5]](#footnote-5), Python library for accessing the Twitter API[[6]](#footnote-6). Twitter provides a large set of filtering parameters to obtain data. This paper initially was focusing on five neighbourhoods in Toronto: (1) The Annex (2) Trinity-Bellwoods (3) Yorkville (4) Regent Park (5) Parkdale. Ideally the neighbourhood names and their variations were to be used as keywords to extract those tweets that contained them as well as well as the geolocation for Toronto.

However, in attempting to do an initial pull from Twitter, the limitation of only having academic access to pull 30 days’ worth of tweets, resulted in few tweets related to Toronto neighbourhoods. On top of that constraint, the context of a global pandemic in the form of the COVID-19 virus[[7]](#footnote-7) had turned the majority of tweets to that subject. Not being able to access the full archive of public Twitter data through Twitter’s Search Tweets API: Full-archive did not allow for searching of tweets over multiple years and to restrict searches to before the COVID-19 pandemic.

Despite these drawbacks, I would still like to demonstrate the potential value TSA can bring when applied in social science contexts such as urban planning. I will use the current context of the COVID-19 epidemic as an example where the same methodology can be applied to Toronto neighbourhood data when pulled from Twitter.

For this example, I have pulled tweets about the Coronavirus and Toronto, Montreal and Vancouver to analyze the polarity of sentiments in the tweets in relation to each of the cities and to compare which cities have either more positive or negative tweets on this topic.

The following code was used to search Twitter:



Figure 1. Sample code for extracting tweets about Vancouver and COVID-19.

The same code was run replacing Vancouver with Toronto, then Montreal to get three CSV files for each city.

The datasets and full code can be found in the following Github repository: <https://github.com/anaedX/toronto-neighbourhoods-sa-initial-code.git>

# Approach

The approach that this paper will take is machine-learning by employing a number of different features to build a classifier that can detect tweets that express sentiment or opinion. It will treat the problem as a binary classification, classifying tweets as either positive or negative in supervised ensembles.

## Step 1: Data Collection

Twitter data was collected using the Tweepy Python library and the parameters described in the Data set section of this paper.

The training data set that is used in this paper is the Sentiment140[[8]](#footnote-8) corpus. The CSV file contains 1,600,000 tweets extracted using the Twitter API. The tweets have been annotated (0 = negative, 4 = positive). For the purpose of this paper and due to time limitations and the computational burden of testing 1.6 million tweets only 10,000 positive tweets and 10,000 negative tweets were extracted from the CSV file to use for training.

## Step 2: Data Preprocessing

Both the Twitter and the training data sets were cleaned using the following methods listed below.

Use Python NLTK[[9]](#footnote-9) python library to:

* + Remove URLs
  + Remove Hashtags
  + Remove Mentions
  + Remove reserved words (RT, FAV)
  + Tokenize the tweets
  + Lemmatize tweets – using WordNetLemmatizer[[10]](#footnote-10)
  + Filter Out Punctuation
  + Filter out Stop Words (and Pipeline)

Use Python emoji[[11]](#footnote-11) to replace emoticons and emoji with its equivalent meaning

## Step 3: Feature Extraction

Scikit-learn[[12]](#footnote-12) software machine learning library for the Python is used for feature representation. The library’s Tf–idf term weighting[[13]](#footnote-13) is used to convert training tweets into numeric representation.

## Step 4: Model Training and Testing

The following base classifiers are used:

* Multinomial Naïve Bayes
* Random Forest
* Support Vector Machines – specifically C-Support Vector Classification
* Logistic Regression

The ensemble classified algorithm proposed by Ankit (2018) was used. The function based on Ankit’s algorithm is shown in the figure 2 below.

10-fold cross validation was used to split and train the data for the model.



Figure 2. Function in python based on Ankit’s (2018) ensemble classifier algorithm.

## Step 5: Evaluation

The classifiers are evaluated using accuracy, precision, recall and F1 scores.

## Step 6: Parameter Tuning

Parameters for the classifiers listed below are tuned using Scikit-learn’s meta-estimator Exhaustive Grid Search[[14]](#footnote-14). GridSearchCV generates candidates from a grid of parameter values specified with the param\_grid parameter.

* C-Support Vector Classification – the best *C*, *gamma* and *kernel* values were predicted using a grid of hyper-parameters that were run to try all of their combinations and those with the best results were generated and used. The parameters were tuned to *c*:1, *gamma:* 1, and *kernel:* rbf.
* Multinomial Naïve Bayes – the best *alpha, n-grams range, IDF usage* and *TF-IDF normalization.* The parameters were tuned to *clf\_\_alpha:* 1*, tfidf\_\_norm:* 12*, tfidf\_\_use\_idf:* False, and *vect\_\_ngram\_range:* (1, 2).
* Logistic Regression – the best *C* and *penalty.* The parameters were tuned to *c*:1 and *penalty*: 12.

For the Random Forest classifier, as random forest is already very good at classification (Meinert, 2019, June 5) and exhaustive grid searches are time consuming, it was decided to keep the default parameter settings.

## Step 7: Prediction

The trained model was applied to each of the Toronto, Montreal and Vancouver twitter data sets to generate predicted tags.

# Results

Figures 3 & 4 compare the performance of Ankit’s (2018) ensemble classifier to the base classifiers listed in [Step 4: Model Training and Testing](#_Step_4:_Model).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Techniques | Accuracy (%) | Precision (%) | Recall (%) | F1 (%) |
| Multinomial Naïve Bayes | 74.49 | 74.70 | 74.49 | 74.43 |
| Random Forest | 67.86 | 68.07 | 67.86 | 67.79 |
| C-Support Vector Classification | 69.66 | 69.74 | 69.66 | 69.63 |
| Logistic Regression | 69.91 | 70.06 | 69.91 | 69.86 |
| Ankit’s Ensemble Classifier | 82.29 | 82.06 | 82.26 | 82.16 |

Figure 3. Cross comparison of the results obtained from base classifiers and Ankit's (2018) ensemble classifier.

Figure 4. Comparison of the base classifier and ensemble classifier evaluations.

The results show that the proposed ensemble classifier, when run on 10, 000 of the extracted Stanford Sentiment 140 corpus tweets, performed more accurately, with better precision, recall and F1 scores than any of the base classifiers.

|  |  |  |
| --- | --- | --- |
| City | Positive Tweets (#) | Negative Tweets (#) |
| Toronto | 1184 | 1209 |
| Montreal | 1400 | 1179 |
| Vancouver | 1348 | 1122 |

Figure 5. Results for Toronto, Montreal and Vancouver data sets after running the ensemble classifier.

Figure 6. Results for Toronto, Vancouver and Montreal after running the ensemble classifier.

The ensemble classifier is run on the Toronto, Montreal and Vancouver data sets with the results indicated in figures 5 & 6. Looking at the data we can see that Toronto’s data set as more tweets that were classified having a negative sentiment than either of the other data sets. Toronto also had a lower number of tweets that were classified as having a positive sentiment overall. If we were looking at this chart and these cities were replaced by tweets about neighbourhood sentiments we could argue that given the data we are seeing Toronto would potentially be seen as the neighbourhood that would need further investigation for revitalization support, given that there is more negative sentiment and less overall positive sentiment tweeted about the area.

# Conclusion

The ensemble classifier performed better than the base classifiers for this report, affirming the results Ankit (2008) had found in their paper *An Ensemble Classification System for Twitter Sentiment Analysis.* Using the ensemble classifier algorithm and the steps that are proposed in this paper, the process can be applied to tweets collected using Twitter’s Search Tweets API: Full-archive when keywords searches for different neighbourhoods are applied. The results from the TSA can be used to then inform key decision makers and urban planners when having to decide which neighbourhoods require support for revitalization over others, especially since funding for these types of projects is finite. Data collection methods using Twitter with NLP sentiment analysis can potentially help to either alleviate some of the burden around the creation and production of civic surveys or can complement the data collected by those means, providing new approaches to evidence gathering and analysis for governments that are seeking them.

# References

Ankit, N. S. (2018). An ensemble classification system for Twitter sentiment analysis. *Procedia Computer Science, 132*, 937-946. <https://doi.org/10.1016/j.procs.2018.05.109>.

Barnaghi, P., Breslin, J. G., & Ghaffari, P. (2016). Opinion mining and sentiment polarity on twitter and correlation between events and sentiment. *IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*, Oxford. <https://doi-org.ezproxy.lib.ryerson.ca/10.1109/BigDataService.2016.36>

Bouazizi, Mondher & Ohtsuki, Tomoaki. (2016). A pattern-based approach for sarcasm detection on Twitter. *IEEE Access*, *4*, 1-1. <https://doi.org/10.1109/ACCESS.2016.2594194>.

Giachanou, Anastasia & Crestani, Fabio. (2016). Like it or not: A survey of Twitter sentiment analysis methods. *ACM Computing Surveys*, *49*(2), 1-41. <https://doi.org/10.1145/2938640>.

Gibbons, J., Malouf, R., Spitzberg, B., Martinez, L., Appleyard, B., Thompson, C., et al. (2019) Twitter-based measures of neighborhood sentiment as predictors of residential population health. *PLoS ONE*, *14*(7): e0219550. <https://doi.org/10.1371/journal.pone.0219550>.

Hu, Y., Deng, C. & Zhou, Z. (2019). A semantic and sentiment analysis on online neighborhood reviews for understanding the perceptions of people toward their living environments. *Annals of the American Association of Geographers*, *109*(4), 1052-1073. <https://doi.org/10.1080/24694452.2018.1535886>.

Meinert, R. (2019, June 5). *Optimizing Hyperparameters in Random Forest Classification.* Medium. <https://towardsdatascience.com/optimizing-hyperparameters-in-random-forest-classification-ec7741f9d3f6>

Nguyen, Q.C., Brunisholz, K.D., Yu, W. et al. (2017). Twitter-derived neighborhood characteristics associated with obesity and diabetes. *Scientific Reports, 7*, 16425. <https://doi.org/10.1038/s41598-017-16573-1>.

Permentier, M., Bolt, G., & van Ham, M. (2011). Determinants of neighbourhood satisfaction and perception of neighbourhood reputation. *Urban Studies*, *48*(5), 977-996. <https://doi.org/10.1177/0042098010367860>.

Saeidi M., Bouchard G., Liakata M. & Riedel S. (2016). Sentihood: targeted aspect-based sentiment analysis dataset for urban neighbourhoods. *Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers*, 1546–56.

Strong Neighbourhood Task Force. (2005). *Strong neighbourhoods – A call to action*. Retrieved from the United Way Greater Toronto website: <https://www.unitedwaygt.org/document.doc?id=61>.

Tuhus-Dubrow, R. (2014, October 21). *Will Twitter revolutionize how cities plan for the future?* Next City. <https://nextcity.org/daily/entry/urban-planning-twitter-social-media-data>.

1. <https://www.edmonton.ca/projects_plans/communities_neighbourhoods/neighbourhood-revitalization.aspx> [↑](#footnote-ref-1)
2. <https://www.saskatoon.ca/engage/2018-civic-surveys-results> [↑](#footnote-ref-2)
3. <https://www.washingtonpost.com/technology/2019/02/07/twitter-reveals-its-daily-active-user-numbers-first-time/> [↑](#footnote-ref-3)
4. <https://sites.tufts.edu/ualab/> [↑](#footnote-ref-4)
5. <https://www.tweepy.org/> [↑](#footnote-ref-5)
6. <https://developer.twitter.com/en/docs> [↑](#footnote-ref-6)
7. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019> [↑](#footnote-ref-7)
8. <http://help.sentiment140.com/for-students> [↑](#footnote-ref-8)
9. <http://www.nltk.org/> [↑](#footnote-ref-9)
10. <https://www.nltk.org/_modules/nltk/stem/wordnet.html> [↑](#footnote-ref-10)
11. <https://pypi.org/project/emoji/> [↑](#footnote-ref-11)
12. <https://scikit-learn.org/stable/modules/classes.html?highlight=feature%20extraction#module-sklearn.feature_extraction> [↑](#footnote-ref-12)
13. <https://scikit-learn.org/stable/modules/feature_extraction.html> [↑](#footnote-ref-13)
14. <https://scikit-learn.org/stable/modules/grid_search.html#exhaustive-grid-search> [↑](#footnote-ref-14)