

# Sentiment Analysis of COVID-19 Social Distancing Tweets

**Jian Sun, Diwen Zhu, Zihao Di**

{j248sun ,d55zhu , z2di}@uwaterloo.ca

University of Waterloo

Waterloo, ON, Canada

## Introduction

This paper describes an artificial intelligence platform for sentiment analysis on Covid-19 related Tweets in Canada and the United States. The Tweets will be classified into three categories: positive, neutral and negative. The goal is to build and train a neural networks classifier to gauge public opinion on social distancing orders in a specific area in real time. This project will also analyze the accuracy of the neural networks algorithm when applied to Tweets of a specific topic such as Covid-19. Having knowledge of public sentiment on Covid-19 could help local authorities improve and adjust their social distancing policies to better suit the needs of the public in a timely manner, which is especially important amid this unprecedented situation.

Twitter is a social networking platform that allows users to post messages called Tweets with a limit of 280 characters. Twitter has two main components: a user timeline that shows Tweets sent by a particular user and a home timeline which displays Tweets of users that a user follows.

As the Covid-19 pandemic continues to develop around the world, governments have started to implement strict social distancing orders in order to stop the spread of the virus. The implementation of these measures, however, are often met with varying degrees of opposition from the public: some have reported they are unable to get essential supplies because of the lack of transportation options while others say the lack of social interactions is taking a heavy toll on their mental health (Breen 2020).

The real-world impact of this project is its ability to allow local authorities, government officials and decision makers in the healthcare industry to better understand the public opinion on the implemented measures in real time. This information can be used to make adjustments and improvements to the measures, which will make social distancing more effective in combating the pandemic. For example, if sentiment in an area is overwhelmingly negative, authorities can immediately investigate and see if social distancing is causing unintended hardship for some. After making changes to their policy, authorities can continue to monitor sentiment levels to see whether it improves, which helps them determine whether these changes are effective.

## Related Work

Since 2001, there has been an explosive growth in sentiment analysis research due to the development of machine learning techniques and the wide availability of datasets (Pang & Lee, 2008). With the rise of microblogging platforms such as Twitter in recent years, researchers have begun to develop better sentence-level analysis techniques instead of reusing those that analyze large opinionated pieces such as product reviews (Kim & Hovy 2004). These generally don't seem to work well with texts that are more expressive and less structured, such as Tweets (Zhang et al. 2011).

Most of the research in the area of sentence-level analysis is either lexicon-based or machine learning-based. The lexicon-based approach uses a precompiled list of opinion words to detect sentiment (Kim & Hovy 2004). However, when applied to sentence-level analysis, this approach can result in low recall (Zhang et al. 2011; Wan & Gao 2015). The machine learning-based approach generally uses supervised learning methods to train classifier models and requires labeled datasets (Zhang et al. 2011; Wan & Gao 2015; Sosa 2017).

One paper from Wen & Gao (2015) proposes an ensemble method that combines five separately trained machine learning models including Naive Bayesian, Bayesian Network, SVM, Decision Tree, and Random Forest. Although this combined approach outperforms each of the five models individually with an accuracy of 84.2%, it is only 0.5% higher than that of the runner-up, C4.5 Decision Tree which has an accuracy of 83.7%. Furthermore, this method uses the majority vote of the five models with equal weighting to make its final decision, which may not be optimal as some of the models are more accurate than others individually.

Another idea proposed by Sosa (2017) combines Long-short Term Memory Neural Networks (LSTM) and Convolutional Neural Networks (CNN). In this combined approach, LSTM encodes the input tokens while maintaining information of all previous tokens, and CNN finds patterns using the richer encoded input. By taking advantage of both methods, the LSTM-CNN achieves an accuracy of 75.2% which is better than either LSTM (72.5%) or CNN (66.7%) individually. The results from a similar approach taken by Severyn & Moschitti (2015) using deep convolutional neural networks also produces similar results.

The approach used in this paper builds on the idea pro-

posed by Sosa (2017). A simplified version of the LSTM model will be trained using some of the structure and parameters presented in Sosa's implementation. This project seeks to achieve comparable accuracy by training the model on a more specific dataset (Covid-19 related) and by tweaking the parameters of Sosa's implementation. The results of the simplified implementation will be compared to Sosa's implementations to determine how well the simplified version works compared to much more sophisticated models.

## Methodology

A simplified version of the LSTM neural network will be built and trained based on the structure and parameters of the Sosa (2017) implementation. Libraries functions from PyTorch, Keras and/or Tensorflow will be used to construct the network. A dataset of 75 million Covid-19 related Tweets will be used to train and test the model. The training dataset will be preprocessed to include an even distribution of positive, neutral and negative Tweets. Labelling of the training data will be done by AWS Comprehend. Accuracy of the model can be measured by comparing the results of the simplified model with the results of running AWS Comprehend on the same test dataset. A manual review of a small subset of the predictions of the test dataset will also be done.

## Results

The accuracy of the simplified model is expected to perform close to the results of the LSTM-CNN model as proposed by Sosa (2017). The simplified model should at least give results better than random guessing.

## References

- Breen, K. 2020. Keeping up social distancing could pose challenge as coronavirus pandemic wears on.
- Kim, S.-M., and Hovy, E. 2004. Determining the sentiment of opinions. proceedings of the 20th international conference on computational linguistics (coling 04). *Association for Computational Linguistics* 1367.
- Lei, Z.; Ghosh, R.; Dekhil, M.; Hsu, M.; and Lu, B. 2011. Combining lexicon-based and learning-based methods for twitter sentiment analysis. *Hewlett-Packard Development Company, L.P.* (2011).
- Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* (2):1?135.
- Severyn, A., and Moschitti, A. 2015. Twitter sentiment analysis with deep convolutional neural networks. *Association for Computing Machinery*.
- Sosa, P. M. 2017. Twitter sentiment analysis using combined lstm-cnn models. *University of California, Santa Barbara*.
- Wan, Y., and Gao, Q. 2015. An ensemble sentiment classification system of twitter data for airline services analysis. *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*.