Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Tracking social media during the COVID-19 pandemic: The case study of lockdown in New York State

Lin Miao\textsuperscript{a,1}, Mark Last\textsuperscript{a}, Marina Litvak\textsuperscript{b}

\textsuperscript{a} Ben-Gurion University of the Negev, P.O.B. 653, Be’er Sheva 8410501, Israel
\textsuperscript{b} Shamoon College of Engineering, 56 Bialik St., Be’er Sheva 8410802, Israel

\section*{ARTICLE INFO}

\textbf{Keywords:}
- Stance detection
- Opinion monitoring
- Social media
- Data distillation
- Government policy

\section*{ABSTRACT}

Facing the COVID-19 pandemic, governments have implemented a wide range of policies to contain the spread of the virus. During the pandemic, large amounts of COVID-19-related tweets emerge every day. Real-time processing of daily tweets may offer insights for monitoring public opinion about intervention measures implemented. In this work, lockdown policy in New York State has been set as a target of public opinion research. This task includes two stages, stance detection and opinion monitoring. For the stance detection stage, we explored several combinations of different text representations and classification algorithms, finding that the combination of Long Short-Term Memory (LSTM) with Global Vectors for Word Representation (GloVe) outperforms others. Due to the shortage of labeled data, we adopted the data distillation method for the training data augmentation. The augmentation of the training data allows to improve the performance of the model with a very small amount of manually-labeled data. After applying the distillation method, the accuracy of the model has been significantly improved. Utilizing the enhanced model, automatically classified tweets are analyzed over time to monitor the public opinion. By exploring the tweets in New York from January 22nd until September 30th, 2020, we show the correlation of public opinion with COVID-19 cases and mortality data, and the effect of government responses on the opinion shift. These results demonstrate the capability of the presented method to effectively and efficiently monitor public opinion during a pandemic.

\section*{1. Introduction}

As many countries are facing unprecedented challenges from COVID-19, the strain on the governments to handle this pandemic and contain the spread of the disease is extreme. Governments have implemented a range of different measures to limit the spread of the COVID-19 virus, safeguard people’s lives and health, and secure adequate health care capacity. COVID-19 control measures require cooperation between government and citizens, as the worldwide impact of the pandemic continues to grow. Hence, it is critical for decision-makers to monitor public opinion about the intervention measures over time, so they can adjust their policy accordingly. During the COVID-19 pandemic, social media has seen a significant increase in use; therefore, the Twitter platform may be an important source of large amounts of data for real-time opinion monitoring. This paper aims at using social media for real-time monitoring of public opinion about intervention measures during the COVID-19 pandemic. To monitor public opinion on social media, stance detection is conducted to identify agreement or disagreement on a certain control measure, such as the lockdown. This task can be successfully handled by supervised learning algorithms, which usually rely on large amounts of annotated data. As such, the limited amount of manually-labeled tweets related to COVID-19 presents a major challenge for real-time public opinion monitoring.

Facing the shortage of annotated stance detection data, we apply a data distillation method for data augmentation. In our previously published work (Miao et al., 2020), we explored several data augmentation methods, showing that with a small manually-labeled dataset, data distillation outperforms other methods on this task. Data distillation is a simple omn-supervised learning method that uses labeled and unlabeled data, together with self-training, to enhance the performance of the model (Radosavovic et al., 2018; Furlanello et al., 2018; Zhang and Sabuncu, 2020). The previous studies have shown that when updated with unseen data in several iterations, data distillation models can reach a consistent improvement.

After applying stance detection to COVID-19-related tweets, we monitor the changes in public attitude (against or in favor of
intervention measures), as expressed in these tweets over time. To more deeply understand these changes, we sought the answers for the following questions:

- How does public opinion about intervention measures change over time during the COVID-19 pandemic?
- How are COVID-19 statistics (e.g., confirmed cases, death cases) related to public opinion?
- How do government intervention measures affect the public opinion?
- What triggers the changes in distribution of negative and positive stance tweets?

Based on these questions, we explore the correlation between public opinion, government policies, and COVID-19 statistics.

The main contributions of our work are summarized as follows:

- We explored several text representations and classification approaches to monitor public attitude towards intervention measures during COVID-19 pandemic.
- We conducted a comprehensive analysis of public opinion about COVID-19 in NY State from January 22nd to September 30th, 2020, to understand its relations with COVID-19 statistics and the intervention measures taken by the government.
- We evaluated the performance of the data distillation method for training data augmentation using only a small set of manually-labeled samples.
- We created and released a new dataset of COVID-19-related tweets, which can be helpful for other researchers.

2. Related work

Real-time information from online social networks has been widely used for various emergency incidents and events. During the H1N1 Influenza pandemic in United States, Signorini et al. (2011) used Twitter data to build a regression model to measure public concerns about that pandemic and to track public sentiment about H1N1, in order to estimate the disease activity level in real time. During the 2016 Zika virus epidemic, Masri et al. (2019) demonstrated the feasibility of utilizing Twitter data for disease surveillance on a national and state (Florida) level. To explore the possibility of detecting influenza outbreaks by analyzing Twitter data, Culotta (2010) correlated the influenza-related tweets with the statistics from the Centers for Disease Control and Prevention (CDC).

During the COVID-19 pandemic, extensive research has been conducted focusing on Twitter data, to analyze public sentiment and responses. To analyze epidemic-related content in social networks, Li et al. (2020) trained a traditional Support Vector Machine (SVM) classifier on manually-labeled data and applied the induced SVM model to label the rest of the data automatically. To understand the temporal sentiment, Wang et al. (2020) used an excon-based model to output the sentiment score of each tweet. They also used a case study format to explore public attitude towards specific measures.

Most of the sentiment analysis studies deal with general public sentiment on some events or during some specific time period. In our work, we are interested in exploring public opinion about intervention measures taken by the NY State government during the COVID-19 outbreak, which is more than just the sentiment. Public opinion is defined as stance (support or against) about a target topic, regardless of whether positive or negative language is used in the text.

2.1. Stance analysis

In terms of stance analysis on Twitter, general approach is to build supervised classifiers using labeled data. Some works employed the Multinomial Naïve Bayes or the SVM model, with Bag-of-Words (BOW) text representation, while others used n-grams as tokens (D’Andrea et al., 2019; Kunneman et al., 2020; Skeppstedt et al., 2017). Compared with D’Andrea et al. (2019), which manually labeled a small training dataset for supervised classification, Kunneman et al. (2020; Skeppstedt et al. (2017) conducted an expensive annotation process to get sufficient labeled data for machine learning classifiers. In contrast, Lukasik et al. (2016) explored the use of Hawkes Processes, taking into account temporal information in addition to text, for stance classification of Twitter rumors. To reduce the manual labeling effort, Rajadesingan and Liu (2014) presented a semi-supervised retweet-based label propagation method, to detect opinion about a certain topic on Twitter. Unlike most of the works, they utilized the retweeting activity of users to obtain the annotations, requiring only a small amount of manually-labeled data. Recently, deep learning methods have become widespread in stance analysis. In the SemEval-2016 stance detection challenge (Mohammad et al., 2016), several deep-learning-based methods were top performers, such as transfer learning based on Recurrent Neural Networks (RNN) in Zarrilla and Marsh (2016), Convolutional neural network (CNN) was used in Wei et al. (2016), and so on. More recently, an attention-based neural ensemble method was developed by Siddiqua et al. (2019) for tweet stance detection. The authors used multi-kernel convolution to extract higher-level features, then combined densely connected Bidirectional Long Short-Term Memory (Bi-LSTM) and nested LSTM models to predict stance. Augustin et al. (2016) implemented a bidirectional conditional LSTM encoding model for stance classification. However, most of the works on stance detection rely on large amounts of manually-labeled data. In our work, we explore a data distillation method for augmenting the training data by leveraging the unlabeled data. In addition, we also utilize deep learning approaches and state-of-the-art text representation models for the stance classification of tweets.

2.2. Data augmentation on text

Data augmentation is critical for supervised learning to tackle the shortage of labeled training data (Dao et al., 2019). Han et al. (2019) leveraged a large domain-specific corpus to fine-tune a language model, which enabled researchers to augment training data by weak supervision for rumor detection. Different from this work, we leverage a small labeled dataset using a distillation method to augment training data. Liu et al. (2020) proposed a text augmentation method using reinforcement learning to guide conditional text generation. For classification of tweets, Sharifrad et al. (2018) conducted data augmentation by using a combination of knowledge graphs to add the related concepts to the original tweets. To boost the performance of text classification, Wei and Zou (2019) proposed to conduct data augmentation using four operations: synonym replacement, random insertion, random swap, and random deletion. This data augmentation method requires to extend the datasets with complex transformations of the original data. In contrast to these works, we aim to explore automatic data augmentation under weak supervision leveraging of unlabeled data, without requiring data transformation or data generation.

2.3. Distillation methods

Hinton et al. (2015) proposed to use knowledge distillation to transfer knowledge from an ensemble of models into a single model that can be trained easily and rapidly. Cho and Haritharan (2019) conducted a comprehensive experimental analysis to explore the possible factors that may influence knowledge distillation.

Li et al. (2017) used data a distillation method for dialogue generation task. Differently, they aimed to increase the specificity of the trained models to generate dialogue, so their data distillation is to remove the training examples from the dataset, which will not contribute to the specificity of the model in each iteration. The idea of data distillation has been adopted in various semi-supervised learning tasks on weakly-labeled data. Radosavovic et al. (2018) trained a model...
on large amounts of labeled data to generate the annotations on unlabeled data; they then retrained the model using generated annotations. Based on data distillation framework, Liu et al. (2019) distilled predictions from a teacher model to guide the learning of student model.

Inspired by the idea of data distillation, we apply it for training data augmentation, using one small dataset of manually-labeled tweets.

3. Methodology

The methodology pipeline is shown in Fig. 1. First, manually-labeled training data is used to train a classification model for stance detection. Then, after the distillation process, the trained model is applied in real-time on incoming Twitter data for downstream opinion monitoring.

3.1. Tweet representation

Text classification depends on the quality of the text representation model. In this work, we experiment with two kinds of text representations. Bag-of-Words (BOW). A vector represents the frequency of each word in a pre-defined dictionary of words. Pre-trained language model representations. The intuition behind pre-trained language models is to train a model based on large corpora and use the resulting representations on other specific tasks in that language. Recently, pre-trained language models have achieved remarkable success, presenting state-of-the-art results in different tasks of natural language processing.

3.2. Tweet classification

After representing manually-labeled data as numerical vectors, we use them to train a supervised classification model. In this work, we consider SVM, LSTM, and fine-tuned Bidirectional Encoder Representations from Transformers (BERT) for the stance classification task. We evaluate and compare between different combinations of text representations and classification models. The best model is used for monitoring the public opinion.

To overcome the shortage of labeled data, we use a data distillation method for training data augmentation, which exploits the large-scale unlabeled data. In our previous work (Miao et al., 2020), we explored, for data augmentation purposes, the distillation method that is adopted from Liu (2019) and Xie et al. (2020).

First, we use a manually-labeled dataset to train a basic teacher model. We then apply the trained model to unlabeled data to get predicted labels. Following that, we train a student model that is initialized with identical architecture and parameters as the teacher model, and apply it to the union of manually and automatically-labeled data. Next, we use the trained student model as a new teacher model. We repeat the process a number of times. During each iteration, only the student model is tested on the validation set. Finally, we adopt the student model, which provides no further performance improvement on validation set.

The intuition behind this approach is that a high-quality teacher model brings up a good student model; the improvement of the student model will reinforce and strengthen the teacher model, reversely. The better the teacher model, the more accurate will be the labels predicted for unlabeled data, leading to better learning for the student model. On the other hand, during each training iteration, the student model evaluates the reliability of the labels predicted by the teacher model, thereby enhancing the teacher model.

The pseudocode of our data augmentation method is shown in Algorithm 1. Note that we applied data distillation only on deep learning models, because multiple configurable hyper-parameters should be optimized during the data distillation process. The pipeline of the distillation method is shown in Fig. 2.

Algorithm 1

Data Distillation Method

1. Input: Labeled Data L, Unlabeled Data U; Output: Trained Model;
2. while U != 0 do
3. Select a subset S from U;
4. Label all samples of S using TM;
5. \(L = L \cup S\);
6. \(U = U - S\);
7. Initialize Student Model (SM), \(SM < TM\);
8. Train Student Model (SM) using L;
9. if SM is better than TM then
10. \(TM = SM\);
11. else
12. Stop
13. end if
14. end while
15. return Trained Teacher Model (TM)

3.3. Monitoring online public opinion

The collected tweets are labeled with our most accurate classification model. To analyze the tweets over time and monitor the public opinion, we focus on the following aspects.

1. Daily tweet numbers. We analyze the changes of daily numbers to observe the public concerns about the target.
2. Daily stance polarity ratios. We track the daily stance polarity numbers and ratios to reveal daily public opinion changes.
3. Government response timeline. To explore the correspondence between the public opinion and government policies, the announcement dates of COVID-19 intervention measures are aligned with the timeline of daily opinion tracing.

4. Daily COVID-19 statistics. We align the daily COVID-19 statistics with the daily analysis of tweets. From global trends and local spikes, we analyze the changes and the trigger factors. We also measure the correlation between COVID-19 statistics and changes in public opinion.

4. Dataset and data annotation

In this work, we built a tweet dataset LockdownTweets related to the lockdown policy in NY State during COVID-19 pandemic. To build LockdownTweets, we selected tweets satisfying our search criteria. These were selected from a large dataset related to COVID-19 and provided by Chen et al. (2020). First, we selected the tweets written in English. Then, we filtered the corpus to obtain only tweets related to NY State and lockdown, using a list of keywords.

Afterwards, the noisy tweets were removed for various reasons, such as:

1. The very short tweet does not contain enough text for opinion identification.
2. The tweet contains "New York," however, it is not related to "the lockdown of New York." For example, "New York" can be part of "The New York Times" or "New York Post."
3. Duplicates.

The retweets represent a big proportion in LockdownTweets, which is around 0.73. We decided to keep retweets in our dataset because they could also reflect the opinion of the users.

To manually label training data for supervised learning, we built annotation guidelines (details are shown in Appendix A.1). Based on the guidelines, three annotators labeled 1098 tweets with 0.843 inter-annotator Cohen’s kappa coefficient agreement. We used majority vote strategy to determine the final labels for nine tweets where our annotators disagreed. Fig. 3 shows some examples of the manually-labeled tweets. Finally, the labeled dataset is split into Labeled-train (254 Against, 216 Support and 263 None), and Labeled-test (126 Against, 108 Support and 131 None). The labeled dataset will be used to train models for stance detection phase. Besides, there are 38,169 Unlabeled tweets dated from January 22nd until September 30th, 2020. These Unlabeled tweets will be labeled by the stance detection model we finally choose for further opinion monitoring.

5. Experiments

Our evaluations address both tasks performed for tracking public opinion: stance detection and opinion monitoring. For stance detection stage, models were trained using manually-labeled training data. In this work, distillation method was applied to optimize the parameters of the trained model for improving. Then, the best stance detection model was applied to unlabeled data for further analysis and monitoring.

5.1. Stance detection

This part of the experiments aim at measuring the quality of the proposed stance classification models and choosing the best model for further monitoring of public opinion regarding lockdown.

---

NYS-related keywords: New York, NYS, NYC, Governor Cuomo. Lockdown-related keywords: lockdown, stay-at-home, Pause, NY on Pause, shelter-at-home, NYPause, stay home.
Expert Systems With Applications 187 (2022) 115797

5

Table 1
Training datasets.

| Dataset         | Against | Favor | None |
|-----------------|---------|-------|------|
| StanceData      | 1395    | 753   | 766  |
| Sentiment140    | 800,000 (positive) | 800,000 (negative) |
| Labeled-train   | 254     | 216   | 263  |

Fig. 4. Architecture of our LSTM model.

(Devlin et al., 2019), to produce dense vector representations for sentences.

Data Description. We consider three related datasets, containing two external sources, for our supervised training experiments. The details of the training datasets are shown in Table 1.

- StanceData is the dataset from SemEval-2016 Task #6 for stance detection in tweets, containing several topics different from the target in our task.
- Sentiment140 is the training data for sentiment analysis.
- Labeled-train is a small dataset we manually labeled from LockdownTweets for this task.

In our previous work (Miao et al., 2020), we tried to build transfer learning models using only external datasets, getting promising results. Instead of using only external datasets, in this work, we incorporated the external datasets and the lockdown-related dataset, under the assumption that the combined datasets may improve the performance. Due to the imbalance of three classes in the datasets we used, we applied Synthetic Minority Oversampling Technique (SMOTE) for oversampling. As an initial attempt, we conducted the experiments on SVM_BOW to see the results with now-combined dataset and SMOTE. The experiments were implemented using sklearn. The results show that SMOTE does not improve the model’s accuracy. Therefore, we decided not to employ this technique with other models. Moreover, the results of using integrated datasets from different sources are not promising. So, in other classification models, we only used Labeled-train for supervised learning.

Evaluated Classification Models. For the stance classification, we applied LSTM, BERT, and SVM models with different text representations. Then, we applied the distillation method on the best model of those, aiming to improve the model with unlabeled data.

The architecture of our applied LSTM model is shown in Fig. 4. The experiments were implemented using Keras. After examining the text lengths in datasets, we set the maximum sequence length to 55. We used Keras to pad digital sequences to a maximum length. A spatial dropout was applied on the embedding layer to reduce overfitting. We used bidirectional Gated Recurrent Unit (GRU), a faster variant of the LSTM architecture. We relied on the Sigmoid activation function, and learned the weights using the Adam optimizer and Cross Entropy loss. We used a typical batch size of 32.

For fine-tuning BERT on our data and task, we used a pre-trained English bert-base-cased model (Devlin et al., 2019), which has 12 transformer layers, 12 self-attention heads, and a hidden size of 768. We froze all layers of BERT model, attached a dense layer, a softmax layer, and trained the new model. The Adam optimizer and Cross Entropy loss were used for training. Empirically, the initial learning rate was set to 0.00001, and the batch size was set to 32.

In this stage, we employed LSTM_GloVe and fine-tuned BERT with the distillation method. In each iteration, 500 new unlabeled tweets from Unlabeled joined the distillation process.

Accuracy Metrics. We evaluated the performance of induced models on Labeled-test, using macro-averaged precision, recall, f1-score, and accuracy. The formulas are shown as follows.

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

\[
F_1\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}
\]

5.2.1. Results and discussion

The results in Tables 2 and 3 did not support the hypothesis that incorporating other datasets can provide better results. From the results, we can see that using only lockdown-related tweets enables us to build a better model in spite of the small dataset. Besides, compared with Sentiment140, StanceData is the dataset from the same domain, showing better performance, but not good enough as Labeled-train. Also, models using SMOTE did not show a clear improvement. As a result, we used only Labeled-train to train the models without applying SMOTE.

We compared between the results produced by LSTM using different dimensions of GloVe. The results of these comparisons can be seen in Table 4. It shows that using 50 dimensions provides the best results. Therefore, we chose to apply 50-dimensional GloVe word vectors for text representation in all our models.

Table 2
Results obtained by SVM.

| Classifier           | P    | R    | F    | Acc  |
|----------------------|------|------|------|------|
| Labeled-train        | 0.51 | 0.50 | 0.50 | 0.56 |
| StanceData           | 0.33 | 0.33 | 0.24 | 0.30 |
| Sentiment140         | 0.20 | 0.35 | 0.25 | 0.29 |
| Labeled-train + StanceData | 0.47 | 0.46 | 0.45 | 0.48 |
| Labeled-train + Sentiment140 | 0.36 | 0.34 | 0.26 | 0.29 |
| Labeled-train + StanceData + Sentiment140 | 0.36 | 0.33 | 0.26 | 0.29 |

Table 3
Results obtained by SVM with SMOTE.

| Classifier           | P    | R    | F    | Acc  |
|----------------------|------|------|------|------|
| Labeled-train        | 0.46 | 0.46 | 0.46 | 0.49 |
| StanceData           | 0.35 | 0.35 | 0.35 | 0.36 |
| Sentiment140         | 0.20 | 0.35 | 0.25 | 0.29 |
| Labeled-train + StanceData | 0.45 | 0.44 | 0.43 | 0.48 |
| Labeled-train + Sentiment140 | 0.35 | 0.34 | 0.22 | 0.44 |
| Labeled-train + StanceData + Sentiment140 | 0.30 | 0.34 | 0.21 | 0.44 |
Table 5 shows the results of different algorithms with different text representations. As for SVM, the model using BOW outperforms others; with regard to LSTM, using GloVe provides the best performance. Notably, models based on BERT did not show better results when compared with others. It is important to note that both are pre-trained language representations, with GloVe trained on Twitter data and BERT on Wikipedia and BookCorpus. We assume that because GloVe was pre-trained on Twitter, it performs better on our data. As can be seen from Table 5, LSTM_GloVe significantly outperforms other supervised models.

Based on our previous attempt of the distillation method for training data augmentation, we employed it on our lockdown data, in order to get an accurate model without extensive manual annotation. We employed LSTM_GloVe with the distillation method, which converged after three iterations. For the comparison, the distillation method was also applied to fine-tuned BERT, which achieved convergence at the second iteration.

Table 6 compares the results of distillation methods with several benchmarks. From the results, we can see that SVM_BOW slightly outperforms the Majority model. It is noteworthy that the distillation method provides significant improvement of the initial teacher models. However, regardless of the improvement from fine-tuned BERT to BERT_Distillation, LSTM_GloVe_Distillation outperforms BERT_Distillation.

Table 7 shows the detailed performance reports of LSTM_GloVe_Distillation’s, with relatively high precision on three classes. Precision was calculated as a ratio of true positive instances to all instances automatically retrieved (as positive) by a model. We are interested in precision more than in recall be- cause we care less about “missing” some positive samples in monitoring than about identifying actual negative samples as false positives.

We chose LSTM_GloVe_Distillation as the most accurate model for classifying the Unlabeled tweets for opinion analysis.

5.2.2. Error analysis

To better understand our model, we analyzed several error cases, and summarized four reasons for misclassification.

1. Ironic expressions lead to inverse polarity.

The two examples are classified as “Favor,” but the true labels are “Against.” The tweets were meant to be ironic, which is not fully understood by the model.

Example1: I personally believe this lockdown is to keep us from having mass shootings while the banking transition occurs.

Example2: @DavidFDodge1 @daktheDUD @NYR Fan Jay @CNN Cuomo forced Nursing homes to take infected patients. NY has MORE cases and MORE deaths than ANY OTHER COUNTRY ON EARTH. NY has MORE deaths in Nursing homes than ANY other state has TOTAL deaths except NJ. Yeah the lockdown worked great

2. Some tweets need more context to infer the stance.

The example is classified as “Against,” but the annotators labeled it as “Favor,” assuming that something “bad” happens “without a lockdown.”

Example: @ogthottie69 @themarkbanker @NYGovCuomo https://t.co/hfJLM2tVeZ this is what happens without a lockdown.

3. Irrelevant to the target.

The example is classified as “Favor,” but the true label is “None.”

This tweet has “lockdown” and “NY” included, but does not express the stance to “lockdown in New York State.” The model is not able to recognize that the tweet does not express the stance on the target.

Example: People don’t want another lockdown in #NYC and keep the curve flat. Wear a mask; it’s that simple. It’s a virus, not a bacteria. #COVID19 #WearAMask

4. Keywords mislead the model.

Example1 is classified as “Against,” but the true label is “Favor.” The phrase “call for an end to the lockdown” leads the model to predict this tweet as “Against,” because the model observed many “Against” instances with “end lockdown.” Example2 is classified as “Favor,” while the true label is “Against.” The phrase “should be on lockdown” plays an important role.

Example 1: Easy for rich people who can work and self isolate in their gated mansions to call for an end to the lockdown. NYC shows us those who die first are the poor, the old, the people living in crowded homes, the essential workers.

Example2: @seanhannity The only people that should be on lockdown are Cuomo and Dblasio

5.3. Opinion monitoring

In this stage, the best stance detection model was applied to all unlabeled tweets for monitoring the public opinion regarding lockdown.

5.4. Daily opinion analysis

After preprocessing and classifying all unlabeled tweets, we counted the daily number of tweets for each class. To monitor the public opinion, the tweets labeled as “Favor” and “Against” were the most important ones. The numbers of “Favor,” “Against,” and the sum of three classes are shown in Fig. 5.

Fig. 5 shows that from February to the middle of March 2020, few tweets expressed concern about lockdown in NY State. Before March 22nd (when Governor Cuomo announced the statewide stay-at-home order), most of the tweets were in favor of lockdown. As the epidemic spread to NY State, however, a significant increase in the number of
lockdown-related tweets was shown. Clearly, the number of Against tweets presented a sharp increase after the statewide lockdown announcement.

5.5. Opinion shift analysis

To smooth the plots of daily numbers, we applied the 7-day moving average, referring to the statistical methods from the Worldometers website. The daily statistics of COVID-19 in NY State were extracted from the COVID-19 in USA dataset.

To investigate how actions taken by the government influence public opinion, we built the Timeline for State Government Response (details are shown in Appendix A.2). We extracted several important events and aligned them with the 7-day moving average of daily positive cases, daily deaths, number of Against tweets, number of Favor tweets, total number of tweets, and the Against to Favor tweets ratio. The results are shown in Fig. 6.

From Fig. 6, we can get a number of interesting findings:

- Public opinion changes along with the COVID-19 statistics. It can be seen that before positive cases were found in NY State, people were indifferent about the lockdown measures. The number of tweets about lockdown went up along with the daily cases and deaths. However, when the number of daily cases and deaths started to decrease after May, the number of lockdown-related tweets was still seeing a big rise. We could understand from our observations that after a long-time lockdown and seeing the drop of deaths and positive cases, people were more willing to stop the lockdown.
- Government measures affect the public opinion. It appears that measures taken by government trigger local peaks of the daily number of tweets in both Favor and Against opinion.
- Both COVID-19 statistics and government measures can trigger the changes in distribution of Against and Favor tweets. We can see that in the early stage before March, the ratios of both Against and Favor tweets stay at a relatively low level. Basically, at that time more Twitter users were indifferent to the lockdown and did not express any polar opinions in their tweets. After that period, however, the ratio of Favor tweets showed a sharp increase when the infected cases were found in NY State. With the announcement of a statewide lockdown order, the ratio of Favor tweets saw a drop, while the ratio of Against tweets began to increase immediately. As far as the daily number of COVID-19 positive diagnoses and death cases climbing is concerned, the ratio of Favor tweets stays relatively high, while after the daily number of positive and death case decreases, the ratio of Favor tweets goes down as well, and tends to remain stable. It can be seen that, when the "Phase 1 of reopening" was implemented, the ratio of Against tweets reached a peak. It demonstrates that some specific government measures have obvious influence on the public opinion.

5.6. Correlation analysis

To analyze the correlation between public opinion and COVID-19 statistics, we calculated Pearson Correlation Coefficients. We reported the p-value to indicate whether the correlation is significantly different from zero. We calculated the ratio of Favor and Against tweets; the correlation was measured after smoothing data with the 7-day moving average. The results are shown in Fig. 7. The scatterplots are used to visualize the correlation coefficient.

The figure shows that the proportion of Against tweets has a strong positive association with total cases, total deaths, and total recovered cases, while the proportion of Favor tweets has a moderate negative association with the same variables. Because the total cases, total deaths, and total recovered cases are all cumulative numbers, these associations can also be explained by the correlation of the Against to Favor tweets ratio with time. Notably, the corresponding value of statistical significance for the correlation coefficient is zero, representing statistically significant results.

6. Discussion

In this work, we aim at using social media to monitor public opinion about intervention measures during the COVID-19 pandemic. Firstly, stance detection is conducted to identify agreement or disagreement on a certain control measure. Secondly, public opinions are monitored relying on the outputs from stance detection. According to our comparative evaluation of several different approaches, the model built by LSTM with GloVe word vectors and data distillation showed the most promising results for stance detection. In contrast to most of the works on stance detection demanding large amount of manually-labeled data,
Fig. 6. Daily statistics of 7-day moving average aligned with government response timeline.
our data augmentation method is more efficient as it extends the data-sets using small amounts of manually-labeled data.

After automatic classification of unlabeled data with the most accurate stance detection model (LSTM_GloVe_Distillation), we analyzed the NY State lockdown-related tweets from January 22nd until September 30th, 2020, to monitor the public opinion. Aligning the results with COVID-19 data and the timeline of important government response measures shows that changing of COVID-19 statistics could cause the opinion shift, and that intervention measures are able to trigger local peaks in the number of tweets. Specifically, we analyzed the correlation coefficient of COVID-19 data and the daily moving averages of two opinion polarities. It can be seen that the ratio of Against tweets has a strong positive correlation with cumulative COVID-19 statistics, while the ratio of Favor tweets has a moderate negative correlation with it.

7. Conclusions

Social media has been widely used for monitoring public opinion about government policies. In this work, we developed and evaluated a pipeline aiming at monitoring public opinion about intervention policy during the COVID-19 pandemic. The proposed pipeline enhances stance detection and the resulting opinion monitoring by efficiently utilizing data distillation to expand the amount of labelled data. The presented case study demonstrates the analysis of public opinion about intervention measures in NY State from January 22nd to September 30th, 2020. We showed that the analysis performed with the introduced pipeline can gain important insights for the decision-makers. It is recommended to use distillation methods for data augmentation when annotated social media data is limited for supervised opinion mining tasks.

8. Limitations and future work

Limitations of this work include the bias towards populations/countries with wide usage of Twitter. A data sample collected from a social media platform does not necessarily represent the entire population. Because of the Twitter API constraints, only small random samples of tweets are accessible. Furthermore, the insufficient amount of annotated data limits the accuracy of this work, even though we obtained statistically significant results using data augmentation. Besides, only English tweets were considered, while discarding the relevant tweets in other languages, which might have some effect on our results.

In the future work, one could explore more advanced methods to improve data augmentation and provide more accurate tweet classification results. Moreover, multilingual capabilities may be implemented for more comprehensive public opinion monitoring. Additionally, time series forecasting models can be utilized for predicting the public opinion trends. The effect of public opinion on the intervention measures effectiveness may be explored as well.
Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A.2. Detailed Government Responses Timeline

Table A8 shows the detailed government responses timeline built according to Wikipedia.¹⁰

### Table A8 Timeline for Policy Announcement.

| Date      | Policy                                                                 |
|-----------|------------------------------------------------------------------------|
| 22-Mar-2020 | Advisory issued ordering nursing homes to admit patients who test positive for the coronavirus and to not allow testing of prospective nursing home patients. |
| 27-Mar-2020 | All schools statewide ordered to remain closed until April 15.          |
| 28-Mar-2020 | All non-essential construction sites ordered to shut down.             |
| 6-Apr-2020 | Statewide stay-at-home order and school closures extended to April 29.  |
| 9-Apr-2020 | List of businesses deemed essential expanded.                          |
| 14-Apr-2020 | All state residents ordered to wear face masks/coverings in public places where social distancing is not possible. |
| 16-Apr-2020 | Statewide stay-at-home order and school closures extended to May 15.   |
| 1-May-2020  | All schools and universities ordered to remain closed for the remainder of the academic year. |
| 7-May-2020  | Statewide four-phase reopening plan is first announced.                |
| 14-May-2020 | State-wide state of emergency extended to June 13.                    |
| 15-May-2020 | Phase 1 of reopening allowed for counties that met qualifications.     |
| 22-Mar-2020 | Four-phase reopening plan is modified to allow non-essential gatherings of 25 people upon entry of Phase 3, and 50 people upon entry of Phase 4. |
| 22-Jun-2020 | NY City meets conditions for Phase 2, allowing the reopening of outdoor dining at restaurants, hair salons and barber shops, offices, real estate firms, in-store retail, vehicle sales, rental repair, services, cleaning services, and commercial building management businesses. |
| 10-Jul-2020 | Malls allowed to open at 25% capacity for regions in Phase 4, with all patrons required to wear masks. |
| 15-Jun-2020 | Phase 1 of reopening allowed for counties that met qualifications.     |
| 17-Jun-2020 | NY City meets conditions for Phase 4, allowing the reopening of outdoor dining at restaurants, hair salons and barber shops, offices, real estate firms, in-store retail, vehicle sales, rental repair, services, cleaning services, and commercial building management businesses. |
| 19-Aug-2020 | Ban on ticketed music events at bars and restaurants.                   |
| 22-Jan-2020 | All NY City schools ordered to close until April 20.                    |
| 10-Jul-2020 | Malls allowed to open at 25% capacity for regions in Phase 4, with all patrons required to wear masks. |
| 22-Jun-2020 | NY City meets conditions for Phase 2, allowing the reopening of outdoor dining at restaurants, hair salons and barber shops, offices, real estate firms, in-store retail, vehicle sales, rental repair, services, cleaning services, and commercial building management businesses. |
| 10-Jul-2020 | Malls allowed to open at 25% capacity for regions in Phase 4, with all patrons required to wear masks. |
| 14-May-2020 | Phase-wide state of emergency extended to June 13.                    |
| 15-May-2020 | Phase 1 of reopening allowed for counties that met qualifications.     |

Appendix A

Appendix A.1. Guideline for Annotation

Twitter users express their opinions in different ways, including supporting or opposing the target explicitly, supporting or opposing some entities related to the target, or by re-tweeting someone else’s supporting or opposing tweet.

**Target:** Lockdown in New York State

**Opinion Labels:** Support, Against, None

Support/Against

It can be inferred from a tweet that the user supports/against the target, if:

- The tweet supports/opposes the target explicitly.
- The tweet supports/opposes someone or something related to the target, from which we could infer the support of/opposition to the target.
- The tweet does not support or oppose anything, but it contains some clues can infer support or opposition.

None

None of the two cases above.

- The tweet has neutral opinion on the target.
- Cannot conclude the opinion of the target from the tweet.

---

References

Augenstein, I., Rocktaeschel, T., Vlachos, A., Bontcheva, K., 2016. Stance detection with bidirectional conditional encoding, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 876–885.

Chen, Emily, Lerman, Kristina, & Ferrara, Emilio (2020). Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. JMIR Public Health and Surveillance, 6(2), e19273. https://doi.org/10.2196/19273

Cho, J. H., & Hariharan, B. (2019). On the efficacy of knowledge distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4794–4802).

Culotta, A., 2010. Towards detecting influenza epidemics by analyzing twitter messages, in: Proceedings of the first workshop on social media analytics, pp. 115–122.

D’Andrea, Eleonora, Ducange, Pietro, Bechini, Alessio, Renda, Alessandro, & Marcelloni, Francesco (2019). Monitoring the public opinion about the vaccination topic from tweets analysis. Expert Systems with Applications, 116, 209–226.

Diao, T., Gu, A., Ratner, A.J., Smith, V., De Sa, C., Ê., C., 2019. A kernel theory of modern data augmentation. Proceedings of machine learning research 97, 1528.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT.

Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., & Anandkumar, A. (2018). Born again models: Leveraging unlabeled data with knowledge distillation. In Proceedings of the 35th International Conference on Machine Learning (PMLR). PMLR.

Han, S., Gao, J., Ciravegna, F., 2019. Data augmentation for rumor detection using context-sensitive neural language model with large-scale credibility corpus.

Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Kunneman, F., Lambrouj, M., Wong, A., Van Den Bosch, A., & Mollema, L. (2020). Monitoring stance towards vaccination in twitter messages. BMC Medical Informatics and Decision Making, 20, 1–14.

Li, J., Monroe, W., Jurafsky, D., 2017. Data distillation for controlling specificity in dialogue generation. arXiv preprint arXiv:1702.06703.

---

¹⁰ https://en.wikipedia.org/wiki/COVID-19_pandemic_in_New_York_(state) #Government_response
