COMPARING RULE-BASED METHODS AND PRE-TRAINED LANGUAGE MODELS TO CLASSIFY FLOOD RELATED TWEETS

A PREPRINT

Cillian Berragan
University of Liverpool
c.berragan@liverpool.ac.uk

Alex Singleton
University of Liverpool
alex.singleton@liverpool.ac.uk

Alessia Calafiore
University of Edinburgh
acalafio@ed.ac.uk

Jeremy Morley
Ordnance Survey
Jeremy.Morley@os.uk

2022-03-29

ABSTRACT

Social media presents a rich source of real-time information provided by individual users in emergency situations. However, due to its unstructured nature and high volume, it is challenging to extract key information from these continuous data streams. This paper compares the ability to identify relevant flood related Tweets between a deep neural classification model known as a transformer, and a simple rule-based classification. Results show that the classification model out-performs the rule-based approach, at the time-cost of labelling and training the model.

***K*eywords** social media • natural language processing • social interaction

1 Introduction

Twitter presents large continuous feed of information regarding emergency events, contributed through individual users, as these events occur. Many emergency events have been studied in relation to Twitter, including hurricanes and floods in the US (Hughes et al. 2014; Kim and Hastak 2018), Paris terror attacks in 2015 (Reilly and Vicari 2021), and UK flooding events (Saravanou et al. 2015; Brouwer et al. 2017).

Extreme weather events have become increasingly common (Kron, Löw, and Kundzewicz 2019), a trend that is expected to continue (Forzieri et al. 2017), meaning there is an increasing demand to predict and understand how natural disasters develop. Tweets have proved useful in complementing and supporting emergency response in many cases, and often the first reports about emergencies on social media often precede those of mainstream media (Perng et al. 2013; Martínez-Rojas, Pardo-Ferreira, and Rubio-Romero 2018; Kim and Hastak 2018; Laylavi, Rajabifard, and Kalantari 2016). It is therefore important to be able to extract flood related Tweets, removing the noise that often comes with social media streams (Ashktorab et al. 2014).

Much of the past work that has used Twitter to study past emergency events has used keywords to identify relevant Tweets (Kryvasheyeu et al. 2016; Brouwer et al. 2017; Morstatter et al. 2013). This however has several issues, keywords are human selected, meaning they require a pre-existing knowledge of the semantics used to describe targeted events. Certain keywords also do not always relate to these emergency events (Sakaki, Okazaki, and Matsuo 2010; Spielhofer et al. 2016), for example a person may be in *'floods of tears'*. Finally, Tweets relevant to emergency events also do not necessarily contain an obvious keyword (*'Cars are floating down the street!'*), and therefore are unable to be detected. More recent work has considered the ability to use machine learning to classify Tweets into those relevant to emergency events, and those that are irrelevant (Imran et al. 2020; Arthur et al. 2018; Sakaki, Okazaki, and

Matsuo 2010; Li et al. 2018). These studies have utilised a variety of methods, building from classical approaches like Naïve Bayes classification (Imran et al. 2013; Li et al. 2018) and Support Vector Machines (SVMs) (Caragea et al. 2011; Sakaki, Okazaki, and Matsuo 2010), while more recent work has considered the emerging prevalence of neural networks in text-based classification (Caragea, Silvescu, and Tapia 2016; de Bruijn et al. 2020; Nguyen et al. 2017). Traditional machine learning methods however rely on the use of feature engineering to determine model input, are unable to preserve word order, and have limited capability to use context, often over-fitting based on features selected (Caragea, Silvescu, and Tapia 2016). Work with neural networks has shown that given pre-trained word embeddings, they have the capability to outperform these methods (Ghafarian and Yazdi 2020; Caragea, Silvescu, and Tapia 2016; Algiriyage and Prasanna 2021).

This work considers the retrospective classification of a selection of Tweets from past flooding events in the United Kingdom, evaluating the effectiveness of a neural classification model called a transformer against a keyword based approach. This work aims to demonstrate the benefits and costs of the use of new sophisticated methods in natural language processing for this task. Further work is expected to build on this, allowing for information extraction from the relevant Tweets to inform first responders, providing more fine-grained information based on the first-hand experience of individuals like specific property damage, or missing persons, allowing social media to complement existing methods used during flood events [muller2015].

2 Methodology

2.1 Data Collection

2.1.1 Flood Data

A historical dataset containing all *Severe Flood Warnings*, *Flood Warnings*, and *Flood Alerts* issued by the UK flood warning system is available through the UK Government under the Open Government Licence. This data was linked with flood zones from the Environment Agency Real Time Flood-Monitoring API. To reduce the volume of flood events being considered, only *Severe Flood Warnings* occurring after 2010 were selected, leaving a total of 314 individual *Severe Flood Warning* events.

2.1.2 Tweets

The Twitter API v2 was used to extract Tweets from the full historic Tweet archive. For each flood warning the query was constructed using several requirements:

- **Time-frame:** 7 days before to 7 days after flood warning
- **Bounds:** Bounding box of the relevant flood area
- **Parameters:** has *geography*, exclude retweets, exclude replies, exclude quotes

Geographic information associated with every Tweet was required due to the decision to use bounding boxes to pre-emptively filter Tweets in areas not subject to flooding. The new Twitter API now uses a combination of factors to associate geographic coordinates with Tweets which overcomes the issues with limited availability of geotags found with many previous studies (Middleton, Middleton, and Modafferi 2014; Carley et al. 2016; Morstatter et al. 2013). Geography associated with a Tweet may now include either *geotags*, *user profile location* or *locations mentioned in Tweet*. The total number of Tweets extracted was 89,864, with an average of 286 Tweets per flood warning. From this corpus, only a random subset of ~2,500 Tweets were considered for training and evaluation, selecting a subset that balances time constraints and mirroring the corpus size of past work (Caragea, Silvescu, and Tapia 2016; Ghafarian and Yazdi 2020).

2.2 Classification

Figure 1 gives an overview of the classification pipeline used, each Tweet was first pre-processed to normalise usernames and web addresses, and hashtags were parsed to extract words (Pota et al. 2020) (Stage 1). The selected ~2,500 Tweets were manually annotated to train the classification model using Doccano (Nakayama et al. 2018), with 20% used for model validation (Stage 2). The validation subset was then used to evaluate model performance in relation to the simple rule-based approach (Stage 3).

The model builds on the established NLP task of sequence classification, taking token sequences ($\mathbf{x} = \{x_0, x_1 \dots x_n\}$), and predicting a single label (y). A pre-trained transformer language model based on the RoBERTa architecture was used as a base, pre-trained using a corpus of 58 million Tweets (Barbieri et al. 2020)[1].

To construct a rule-based approach for evaluation against this model, Tweets from the validation subset that included a selection of 456 keywords provided by (Saravanou et al. 2015) were labelled as being flood related (*FLOOD*), while all Tweets that did not contain this selection of keywords were labelled as *NOT_FLOOD*.

[1]Available on the Huggingface Model Hub (Wolf et al. 2020)

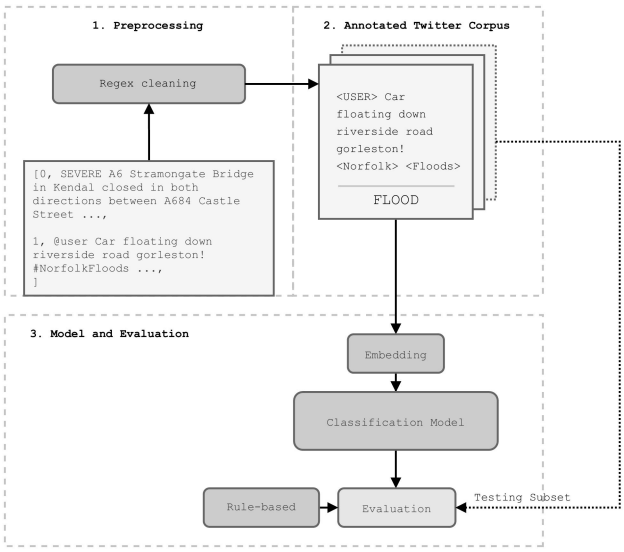

Figure 1: Overview of the model processing pipeline.

For comparative evaluation, the F_1 metric was used, which takes the harmonic mean of the precision and recall, meaning class imbalance is accounted for. To qualitatively assess the performance of the transformer model, *attributions*[2] for each word in a few selected Tweets were visualised to identify the ability of the model to capture information relevant to flood events, without having to explicitly be fed in keywords (Sundararajan, Taly, and Yan 2017).

Overall the classification model out-performed the rule-based method on the validation subset, achieving an F_1 score of 0.938, compared with 0.814 for the rule-base approach. There is both a lower recall for the rule-based model (0.905 compared with 0.952), and a lower precision (0.952 compared with 0.988).

Figure 2 explores the decisions made by the transformer model, using four example Tweets to demonstrate the *attribution* given to each token when assigning a label. Figure **??** (A) first gives an example Tweet that is correctly identified as being flood related by the transformer, but does not contain any selected flood related keywords. In this example three keywords are highlighted as important by the model for its correct classification *gravel, river* and *wier*. This suggests that the model is able to infer from context that these words relate to flooding, rather than having to be explicitly told through feature engineering or keywords.

On Figure 2 (B), an example is chosen where the model was able to correctly identify the Tweet as being unrelated to flooding, but contains the keyword *lightning* meaning the rule-based method incorrectly identified it as flood related. Several keywords again appear important for this correct classification, *finally* which is unlikely to appear in Tweets relevant to floods, in addition to *apples* and *ipad pro*, both of which likely appear relatively frequently on Twitter, but rarely in flood related contexts.

The final two sub-figures give examples where the model gives incorrect classifications, but the rule-based method does not. Figure 2 (C) shows that while the model realises that *raining* is a word positively associated with flooding, the rest of the sentence implies that the overall Tweet is likely not in reference to a flooding event. This example reflects a potential issue with selecting a broad annotation scheme, which considered mentions of weather that may relate to flooding events to be a positive match. A Tweet like this is relatively borderline, even for human annotation, meaning it is unsurprising that the model struggles to make a correct decision. This issue is also reflected in Figure 2 (D), the words *tide, mark* and *kent* are all identified as flood related words, which is likely true and the label reflects an issue with human annotation.

3 Discussion

While the transformer-based classification model outperforms a rule-based approach, they present different benefits and costs. Supervised classification through a neural network relies heavily on a suitable amount of high quality labelled data, which presents an initial time-cost. Keyword selection is comparatively straightforward, and does not rely on a pre-existing corpus of relevant text. The training and inference for the transformer model also costs both time

[2]https://github.com/cdpierse/transformers-interpret

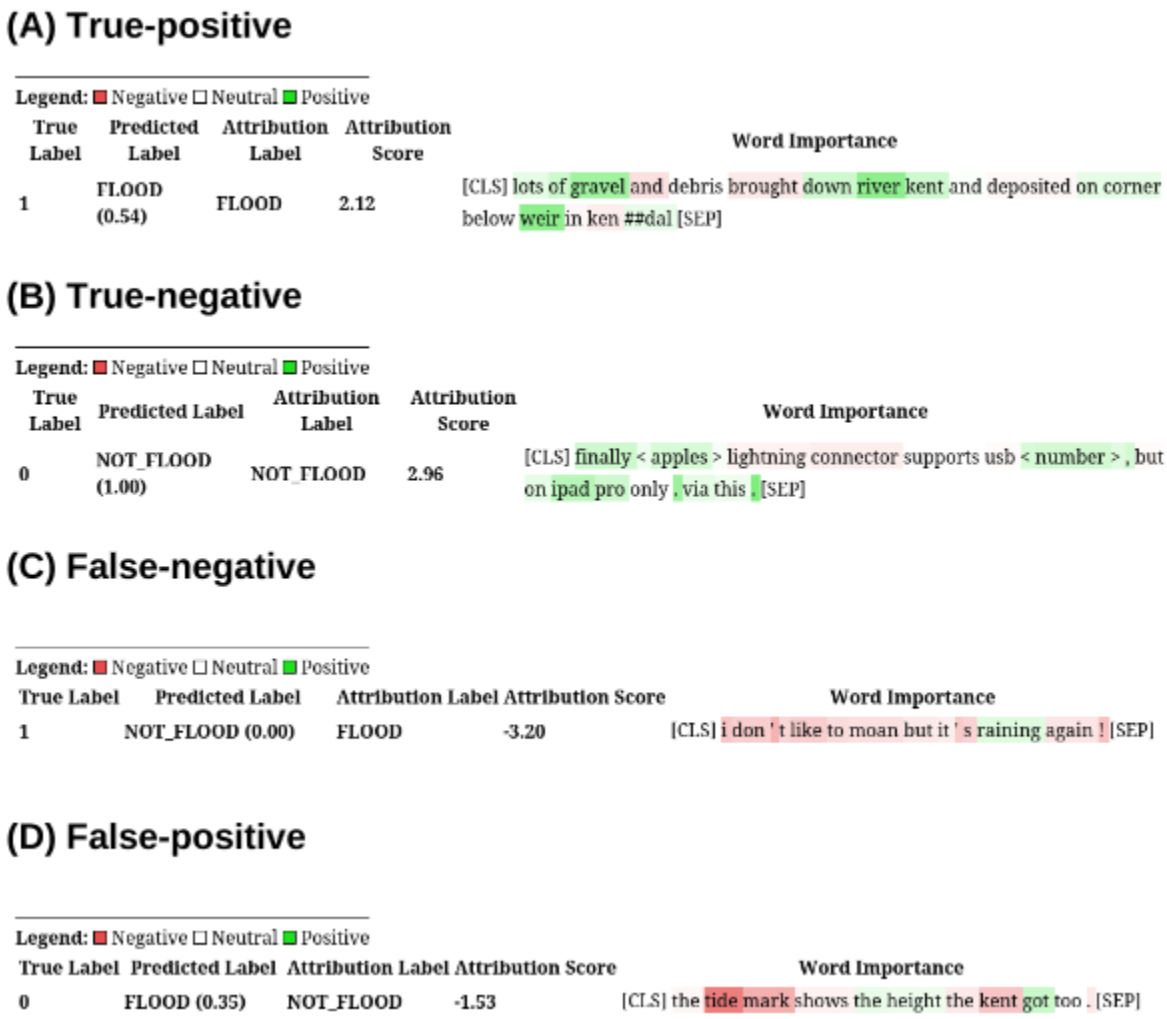

Figure 2: Attribution levels for selected Tweets classified by the transformer model. Attribution label indicates the human annotated label, predicted label shows assigned label with confidence values. Positive attributions dictate the importance of a feature in the given label prediction.

and resources, while keyword selection may be applied directly during the extraction of Tweets through the Twitter API.

Keywords however are inherently subjective, as demonstrated by past work which found varying selections of keywords to be appropriate the classification of flood related Tweets (Spielhofer et al. 2016; Arthur et al. 2018; Saravanou et al. 2015). Constructing a labelled corpus a broad binary classification of Tweets to train a supervised model is less subjective, as the model itself may use the context provided through the training data to independently learn how to approach the classifications. This increased complexity means, through higher recall and precision, the model approach retrieves more relevant information, while ignoring a higher proportion of irrelevant information.

The complexity of the transformer architecture itself also presents improvements over past machine learning methods, as word order is preserved, and the pre-trained word embeddings mean no *ad hoc* feature engineering is required, which may have contributed to some bias and over-fitting in past work (Caragea, Silvescu, and Tapia 2016). Appropriate use of semantic context is a particular benefit of the transformer architecture over simpler deep learning methods, notable on Figure 2 (B), which indicates that while lightning is likely considered by the model in most contexts to be associated with floods, the model is able to consider this instance independently, understanding that in this context the word *'lightning'* is not weather related. Further work should consider using the entire corpus of 89,864 Tweets extracted relating to UK flood events to train a more robust model.

Algiriyage, Nilani, and Raj Prasanna. 2021. "Identifying Disaster-related Tweets: A Large-Scale Detection Model Comparison," 13.

Arthur, Rudy, Chris A. Boulton, Humphrey Shotton, and Hywel T. P. Williams. 2018. "Social Sensing of Floods in the UK." Edited by Guy J-P. Schumann. *PLOS ONE* 13 (1): e0189327. https://doi.org/10.1371/journal.pone.0189327.

Ashktorab, Zahra, Christopher Brown, Manojit Nandi, and Aron Culotta. 2014. "Tweedr: Mining Twitter to Inform," 5.

Barbieri, Francesco, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification." *arXiv:2010.12421 [Cs]*, October. https://arxiv.org/abs/2010.12421.

Brouwer, Tom, Dirk Eilander, Arnejan van Loenen, Martijn J. Booij, Kathelijne M. Wijnberg, Jan S. Verkade, and Jurjen Wagemaker. 2017. "Probabilistic Flood Extent Estimates from Social Media Flood Observations." *Natural Hazards and Earth System Sciences* 17 (5): 735–47. https://doi.org/10.5194/nhess-17-735-2017.

Caragea, Cornelia, Nathan McNeese, Anuj Jaiswal, Greg Traylor, Hyun-Woo Kim, Prasenjit Mitra, Dinghao Wu, et al. 2011. "Classifying Text Messages for the Haiti Earthquake," 10.

Caragea, Cornelia, Adrian Silvescu, and Andrea H Tapia. 2016. "Identifying Informative Messages in Disaster Events Using Convolutional Neural Networks," 8.

Carley, Kathleen M., Momin Malik, Peter M. Landwehr, Jürgen Pfeffer, and Michael Kowalchuck. 2016. "Crowd Sourcing Disaster Management: The Complex Nature of Twitter Usage in Padang Indonesia." *Safety Science* 90 (December): 48–61. https://doi.org/10.1016/j.ssci.2016.04.002.

de Bruijn, Jens A., Hans de Moel, Albrecht H. Weerts, Marleen C. de Ruiter, Erkan Basar, Dirk Eilander, and Jeroen C. J. H. Aerts. 2020. "Improving the Classification of Flood Tweets with Contextual Hydrological Information in a Multimodal Neural Network." *Computers & Geosciences* 140 (July): 104485. https://doi.org/10.1016/j.cageo.2020.104485.

Forzieri, Giovanni, Alessandro Cescatti, Filipe Batista e Silva, and Luc Feyen. 2017. "Increasing Risk over Time of Weather-Related Hazards to the European Population: A Data-Driven Prognostic Study." *The Lancet Planetary Health* 1 (5): e200–208. https://doi.org/10.1016/s2542-5196(17)30082-7.

Ghafarian, Seyed Hossein, and Hadi Sadoghi Yazdi. 2020. "Identifying Crisis-Related Informative Tweets Using Learning on Distributions." *Information Processing & Management* 57 (2): 102145. https://doi.org/10.1016/j.ipm.2019.102145.

Hughes, Amanda L., Lise A. A. St. Denis, Leysia Palen, and Kenneth M. Anderson. 2014. "Online Public Communications by Police & Fire Services During the 2012 Hurricane Sandy." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1505–14. Toronto Ontario Canada: ACM. https://doi.org/10.1145/2556288.2557227.

Imran, Muhammad, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. 2013. "Extracting Information Nuggets from Disaster- Related Messages in Social Media," 10.

Imran, Muhammad, Ferda Ofli, Doina Caragea, and Antonio Torralba. 2020. "Using AI and Social Media Multimodal Content for Disaster Response and Management: Opportunities, Challenges, and Future Directions." *Information Processing & Management* 57 (5): 102261. https://doi.org/10.1016/j.ipm.2020.102261.

Kim, Jooho, and Makarand Hastak. 2018. "Social Network Analysis: Characteristics of Online Social Networks After a Disaster." *International Journal of Information Management* 38 (1): 86–96. https://doi.org/10.1016/j.ijinfomgt.2017.08.003.

Kron, Wolfgang, Petra Löw, and Zbigniew W. Kundzewicz. 2019. "Changes in Risk of Extreme Weather Events in Europe." *Environmental Science & Policy* 100 (October): 74–83. https://doi.org/10.1016/j.envsci.2019.06.007.

Kryvasheyeu, Yury, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, and Manuel Cebrian. 2016. "Rapid Assessment of Disaster Damage Using Social Media Activity." *Science Advances* 2 (3): e1500779. https://doi.org/10.1126/sciadv.1500779.

Laylavi, Farhad, Abbas Rajabifard, and Mohsen Kalantari. 2016. "A Multi-Element Approach to Location Inference of Twitter: A Case for Emergency Response." *ISPRS International Journal of Geo-Information* 5 (5): 56. https://doi.org/10.3390/ijgi5050056.

Li, Lisha, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. 2018. "Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization," 52.

Martínez-Rojas, María, María del Carmen Pardo-Ferreira, and Juan Carlos Rubio-Romero. 2018. "Twitter as a Tool for the Management and Analysis of Emergency Situations: A Systematic Literature Review." *International Journal of Information Management* 43 (December): 196–208. https://doi.org/10.1016/j.ijinfomgt.2018.07.008.

Middleton, Stuart E., Lee Middleton, and Stefano Modafferi. 2014. "Real-Time Crisis Mapping of Natural Disasters Using Social Media." *IEEE Intelligent Systems* 29 (2): 9–17. https://doi.org/10.1109/mis.2013.126.

Morstatter, Fred, Juergen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose," 9.

Nakayama, Hiroki, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. "Doccano: Text Annotation for Humans."

Nguyen, Dat, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. 2017. "Robust Classification of Crisis-Related Data on Social Networks Using Convolutional Neural Networks." *Proceedings of the International AAAI Conference on Web and Social Media* 11 (1): 632–35. https://doi.org/10.1609/icwsm.v11i1.14950.

Perng, Sung-Yueh, Monika Büscher, Lisa Wood, Ragnhild Halvorsrud, Michael Stiso, Leonardo Ramirez, and Amro Al-Akkad. 2013. "Peripheral Response: Microblogging During the 22/7/2011 Norway Attacks." *International Journal of Information Systems for Crisis Response and Management* 5 (1): 41–57. https://doi.org/10.4018/jiscrm.2013010103.

Pota, Marco, Mirko Ventura, Rosario Catelli, and Massimo Esposito. 2020. "An Effective BERT-Based Pipeline for Twitter Sentiment Analysis: A Case Study in Italian." *Sensors* 21 (1): 133. https://doi.org/10.3390/s21010133.

Reilly, Paul, and Stefania Vicari. 2021. "Organizational Hashtags During Times of Crisis: Analyzing the Broadcasting and Gatekeeping Dynamics of #PorteOuverte During the November 2015 Paris Terror Attacks." *Social Media + Society* 7 (1): 205630512199578. https://doi.org/10.1177/2056305121995788.

Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. 2010. "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors," 10. https://doi.org/10.1145/1772690.1772777.

Saravanou, Antonia, George Valkanas, Dimitrios Gunopulos, and Gennady Andrienko. 2015. "Twitter Floods When It Rains: A Case Study of the UK Floods in Early 2014." In *Proceedings of the 24th International Conference on World Wide Web*, 1233–38. Florence Italy: ACM. https://doi.org/10.1145/2740908.2741730.

Spielhofer, Thomas, Reynold Greenlaw, Deborah Markham, and Anna Hahne. 2016. "Data Mining Twitter During the UK Floods: Investigating the Potential Use of Social Media in Emergency Management." In *2016 3rd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, 1–6. Vienna, Austria: IEEE. https://doi.org/10.1109/ict-dm.2016.7857213.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. 2017. "Axiomatic Attribution for Deep Networks." *arXiv:1703.01365 [Cs]*, June. https://arxiv.org/abs/1703.01365.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, et al. 2020. "Transformers: State-of-the-art Natural Language Processing." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.