Data Fusion

Methodology for Detecting Social Media Content about Unemployment

Samuel Fraiberger, Nir Grinberg, David Lazer

June 2019

OBJECTIVE

This part of the project focuses on the detection, in a variety of languages, of social media posts that are related to individuals' employment, including five categories of interest: (1) losing a job to become unemployed, (2) currently being unemployed, (3) restoring employment, (4) searching for a job, and (5) advertising an existing position.

PROBLEM STATEMENT

The technical challenge is to obtain a high precision, high recall classifier for the retrieval of textual information that is rare. Because posts about unemployment in social media are likely to be rare, it is inefficient and prohibitively expensive to obtain labels for a random sample of posts. Thus, we must resort to heuristics that seek to improve recall while preserving high precision. This problem naturally fits into the formulation of *Active Learning* in machine learning, where the learning model is actively involved and iteratively used to select examples for further labeling. At each iteration, the model will select a batch of examples for labeling, use crowdsourcing to obtain labels, and train a new model using all available data.

METHODOLOGY

One of the longstanding and robust methodologies for Active Learning is Uncertainty Sampling [2]. At its heart, Uncertainty Sampling choses to label, at each iteration, instances that the model is most uncertain about. For example, using a logistic regression model for binary text classification, the method would choose to label instances that have predicted probability closest to 0.5 (the threshold separating the two classes). Furthermore, constructing batches for labeling using uncertainty sampling has been

shown to produce comparable results to other, more recent methods, including Adversarial Learning, Core Set approximation, and others. Based on these results, we will populate the batches for labeling in each iteration using Uncertainty Sampling.

The Uncertainty Sampling algorithm requires a probabilistic classification model. We will use a pre-trained multi-lingual BERT model and fine-tune its final classification layer [1]. The instances closest to the classification thresholds would be used to obtain a new batch of labels.

The first batch for labeling needs to be sampled differently because a classification model is not yet available and because labeling a random set of posts is likely to yield very little signal for our categories of interest. Thus, we need an unsupervised approach to over-sample the positive class. We use a small set *keywords*, *target sentences*, and the pre-trained model of BERT to construct a stratified sample of posts for the first labeling batch. Notice that further iterations will use a different, fine-tuned (supervised) BERT model to identify candidates for labeling, and will be free to explore the space further. The sole purpose of the stratified first batch is to provide sufficient signal for the classifier to learn the categories of interest, which are rare otherwise.

The stratified first batch would be constructed as follows. Let P and N be the complementary sets of posts in our data that contain any or none of the keywords in K, respectively. In addition, let T be the set of target sentences, a set of short example sentences that neatly fit our categories of interest (see Appendix A). For every keyword and target sentence pair (k, t), we will use pre-trained BERT to rank the subset of posts in P that contain keyword k based on their similarity to target sentence t (see appendix B). The top 100 ranked posts will be inserted into the batch. In order to reduce the sensitivity to target sentences, we will add to the batch a random sample of 100 posts that contain k, regardless of their ranking. Similarly, to avoid keyword-sensitivity, we will add the top 100 posts from N that are most similar to target sentence t using BERT's similarity, and by construction do not have any of the keywords in K. Using a set of seven keywords and ten target sentences, we will label 7,700 posts from P and a thousand posts from N, resulting in an initial batch size 8,700 for labeling.

The labeling of posts will be obtained via crowdsourcing. Two native-language speakers would annotate each post and a third annotator will be invited in case of disagreement.

CROWDSOURCING

Recruiting: We would recruit annotators to label posts using Prolific.ac based on reported languages they are fluent in, and when possible country of residence.

Task: Annotators will be given instructions to read the content of social media posts and asked to answer a series of questions about it. See Appendix C for the specific categories of interest and the particular survey items.

REFERENCES

[1] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. (2018).

[2] Lewis, D.D. and Gale, W.A. 1994. A Sequential Algorithm for Training Text Classifiers. *SIGIR '94* (1994), 3–12.

[3] Liang, P. 2005. *Semi-supervised learning for natural language*. Massachusetts Institute of Technology.

[4] Meng, Y., Shen, J., Zhang, C. and Han, J. 2018. Weakly-Supervised Neural Text Classification. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (New York, NY, USA, 2018), 983–992.

APPENDIX A

Target sentences

'I lost my job today'
'I was fired earlier this week'

'Now I am unemployed'
'I am currently not working'

'I am searching for a new position'
'Anyone hiring?'

'I got hired today'
'I recently started working at my new job'

'Here is a job opportunity you might be interested in'
'Looking for a new position?'

Keywords

job

work

quit

position

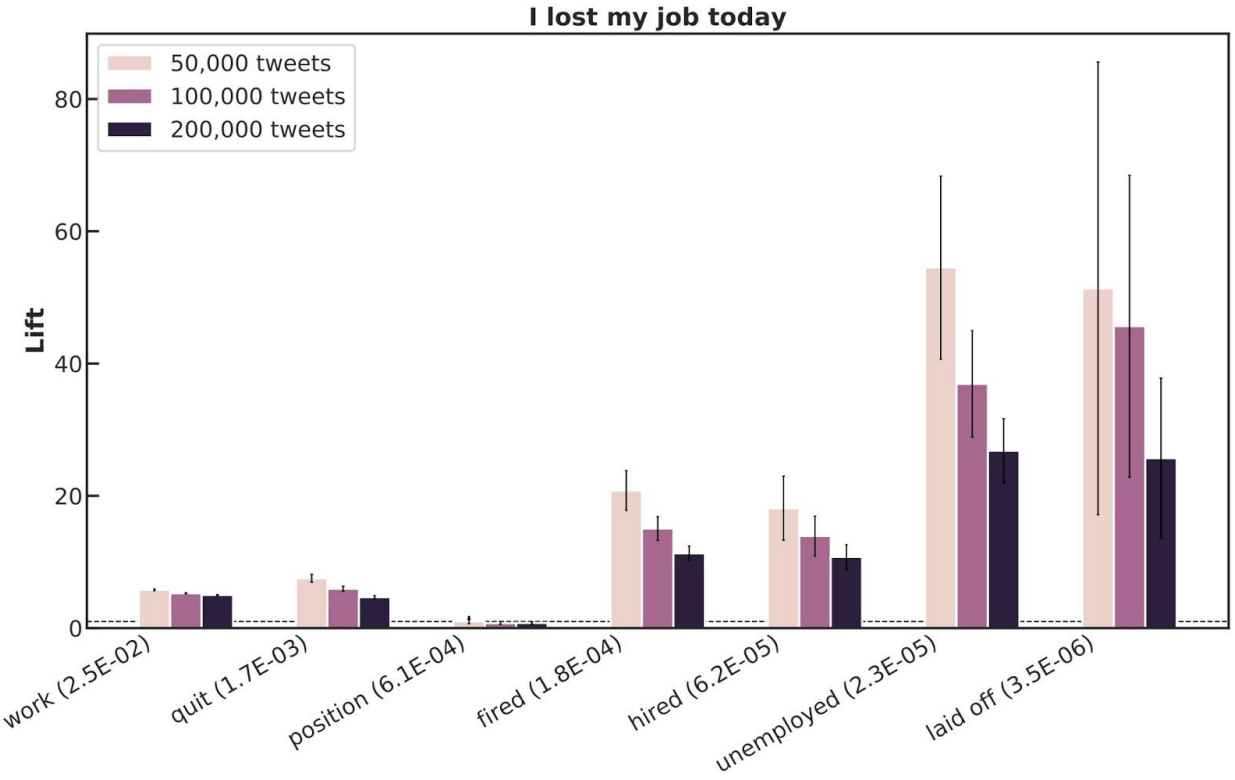
fired

hired

unemployed

laid off

APPENDIX B



Note: Using a random sample of 10 million English tweets, this figure shows the probability of occurrence of each selected keyword in a tweet, after having ranked each tweet by its

similarity to the target sentence “I lost my job today”. Similarities are measured using the pre-trained BERT algorithm. Probabilities of occurrences are calculated using the k tweets that are the most similar to the target sentence, for k in {50,000; 100,000; 200,000}. We then report these probabilities relative to the average probability of occurrence of each keyword in our sample (“lift”). Lift factors vary from 1 to 55 across keywords whose unconditional probability of occurrence (reported between parentheses) vary by several orders of magnitude. Similar results are obtained across target sentences. This figure indicates that our first batch of labeled tweets is likely to contain relevant keywords that could have been missed when choosing our keywords list, allowing to reduce the sensitivity of our algorithm to the keywords list.

APPENDIX C

Survey items

Annotator will be required to answer the following list of questions for each post:

1. Does this tweet indicate that the user recently lost or left a job? [Answer: YES/NO] Examples: I lost my job today; I was fired earlier this week; I recently got laid off; I just quit my job.
    a. [Conditional on a Yes answer the following will be shown] Does this tweet indicate that the user had multiple jobs? [Answer: YES/NO] Examples: Can no longer do this multiple jobs thing, I’m quitting; Today I got fired, luckily I still have my other jobs.
2. Does this tweet indicate that the user is currently unemployed? [Answer: YES/NO] Examples: Now I am unemployed; I am currently not working.
3. Does this tweet indicate that the user is searching for a job? [Answer: YES/NO] Examples: I am looking for a job; I am searching for a new position; Anyone hiring?
    a. [Conditional on a Yes answer the following will be shown] Does this tweet indicate that the user is looking to have multiple jobs? [Answer: YES/NO] Examples: Need to find a second job, anyone hiring? Not making enough money, need to find a night job.
4. Does this tweet indicate that the user was recently hired? [Answer: YES/NO] Examples: I just found a job; I got hired today; I started working at my new position on Monday; Can’t wait to start my new job!
    a. [Conditional on a Yes answer the following will be shown] Does this tweet indicate that the user has multiple jobs? [Answer: YES/NO] Examples: Just

started my second job today! Tomorrow I will start my new job. I hope juggling this with my other positions is worth it.

5. Does this tweet contain a job offer? [Answer: YES/NO] Examples: Looking for a new position? Here is a job opportunity you might be interested in.