Completely Unsupervised Opinion Mining in online (professional groups') discussions

Jafar Mansouri[1&2], Fabrice Cavarretta[2], Wassim Swaileh[1], Dimitris Kotzinos[1]

[1] ETIS Lab, UMR 8051, CY Cergy Paris University, ENSEA, CNRS; Pontoise, France
[2] ESSEC Business School, Cergy, Paris France

The explosion of online discussions in different types of social media provides us with a large corpus of continuous text exchanges over a variety of different topics. Trying to automatically extract and mine those opinions brings up two distinct but highly related problems: (i) the need of identifying relevant posts (i.e., classify posts as relevant or not to the subject of interest) and (ii) extract opinions from those posts and subsequently reclassify them in different classes in order to assess e.g. the importance of the different subjects/opinions. Many works rely on supervised classification methods [1], which means that an already labeled dataset has been provided to the method and used to train a Machine Learning classifier. These methods suffer from inherent bias, i.e., the quality classification can be biased by the labeling. For various types of studies, this prohibits the use of supervised methods. In this paper, we propose the completely unsupervised extraction of opinions of a specific professional group based on the posts on the social media platform Twitter. We want to focus on opinions related to the professional activity of the group, which was the group of entrepreneurs but the methodology described can be applied to any professional group. We used the Tweepy API [4] to collect tweets in the English language and we defined the groups of interest based on the self-descriptions of the users on their profiles (self-labeled as "entrepreneurs"). We collected about 47M tweets from about 24K users/entrepreneurs and around 53M tweets from 38K users/general public (with the requirement not to have the above keywords on their profile), dating from September 2020 and back. The public set plays the role of a control group, representing the topics of the general discussions.

The proposed method eliminates the need of a pre-labeled training set for classifying relevant and not tweets and allows us to work in an unsupervised manner and avoid bias. We rely on the fact that usually specific words or combinations of words can be used to discriminate between two sets of texts when they appear frequently in one set of texts and not frequently in the other set of texts. So, for each set of tweets for the entrepreneurs (ENT) and public (PUB), we find words and combinations of two-words in tweets and their frequencies. Here, frequency means how many *users* in the ENT set and respectively in the PUB set, have used one word or any combination of two words in their tweets. For each user, each word or combination is just counted once. Additionally, for each set, we calculate weights: $w(word_{ent_x}) = (freq_{ent_x}/N_{ent}) - (freq_{public_x}/N_{public})$, in which $freq_{ent_x}$ is frequency of word x in the ENT set, $freq_{public_x}$ is frequency of word x in the PUB set, N_{ent} and N_{public} are the number of users in ent and public, respectively. We do the same procedure for combination of two-words with the difference that we multiply weights by 2 since a combination of two-words is more informative than one word. Note that first we do tweet cleaning, i.e., we remove common stop-words, emojis, targets and links and numbers and punctuations. For each tweet, we find all single words and combinations of two-words; add their weights; divide each sum by the number of words and combinations to find a normalized score for that tweet: $\sum_{i=w_1}^{w_N} w_i /N + \sum_{j=c_1}^{c_M} c_j /M$. We sort tweets in the ENT set based on their scores to find the most relevant tweets. Then we select the first *K* non-repetitive tweets (positive tweets) from each set for training the neural network; where *K* is a parameter that can be specified by the user depending on the required quality and the size of the dataset to be classified. To find all relevant tweets in each set, binary classification by a convolutional neural network (CNN) [2] is used. The negative samples for the training are chosen randomly from the public dataset since it is assumed that most of the tweets in the public dataset are largely irrelevant to the professional group discussions. Although there might be some relevant tweets in the public dataset, their number is assumed to be very low and given the random selection their influence on the final results is very minimal. The neural network then learns from the seed tweets and can classify as relevant or not the corpus. Finally, we use neural network-based methods [3] to extract and classify opinions on a dynamic number of classes based on the relevant tweets that have been identified in the first phase of the work.

References

1. Meddeb I., Lavandier C. and Kotzinos D. (2020) "Using Twitter Streams for Opinion Mining: a case study on Airport Noise", Springer Series "Communications in Computer and Information Science" Volume 1197, 2020, Pages 145-160.

2. Yan, Y., Li, W., Chen, G. and Liu, W., (2020). An improved text classification method based on convolutional neural networks. In 2020 International Conference on Control, Robotics and Intelligent System (pp. 185-190).
3. Jianqiang, Z., Xiaolin, G., Xuejun, Z.: Deep convolutional neural networks for twitter sentiment analysis. IEEE Access 6, 23253–23260 (2018). DOI 10.1109/ACCESS.2017.2776930
4. TweePy API, https://www.tweepy.org