Project 1: Data Collection and Annotation of Tweets

Messages shared over social, particularly Twitter, represent powerful source of information for all kinds of data analysis. Many tasks such as sentiment/severity analysis or event detection require two main steps: (1) the collection of data relevant for the task and (2) the typically manual annotation of the data to serve as input for machine learning algorithms. Both steps are not trivial but very important since they highly affect all subsequent outcomes: bad data + bad annotations = bad results. In general, in both steps, quantity and quality matters.

**Crawling Relevant Datasets.** Most basically, the collected data need to match a current task. The optimal solution would be to use a completely arbitrary set of tweets and identify "enough" positive ones; as well as an equally large set of negative tweets. The problem with this approach is that we're often interested in events that are (hopefully) very rare, making the task of collecting a sufficient number of relevant tweets very labor and time-consuming. In practice, in most cases, a more targeted data collection is needed, to increase the likelihood that the dataset already contains a lot of positive/relevant tweets. However, the notion of relevance often clashes with the to collect the data. For example, Twitter allows to search by keyword but not by topic. Furthermore, while Twitter allows search for tweets based on location, most tweets are not geo-tagged.

**Annotating Tweets.** Manually annotating or labeling large sets of tweets can be rather time consuming. Therefore the task annotating or labeling should be as easy and quick as possible. Another challenge is that choosing an annotation, label, class, etc. can be highly subjective. This is particularly true in case of sentiment or severity analysis. But also identifying entities for Named Entity Recognition (NER) is often not straightforward. For example, given the simple tweet "I'm going to Starbucks.", Starbucks can be viewed as company/organization or location (whereas, on the other hand, the latter does not refer to specific (geo) location).

Apart from the more technical challenges, the task of annotating tweets also includes the question of how to incentivize users to actually perform the task. Again, its not only a question about quantity but users also need to have an incentive to be precise when annotating. While students or researchers directly benefiting from the dataset have an intrinsic incentive to ensure a good quality of the annotations, a wider user group might require additional motivations. For example, In Amazon Mechanical Turk (MTurk), workers are getting paid. Good experiments are designed in a way that the payment reflects the quantity as well as the quality of the performed tasks. Measuring the quantity, e.g., in terms of annotated tweets, is typically pretty straightforward. In general, this is less so regarding measuring the quality. Here, the standard approach is to how close a user's annotation is to the consensus.