==========
D A T A S E T / D A T A B A S E D E S C R I P T I O N
==========

(template based on https://arxiv.org/abs/1803.09010)

* Name of the dataset/database: **“Twitter opinion on bonus intentions KLM-Air France while receiving government support”**

==========
1. MOTIVATION
==========

1.1 For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created in response to KLM-Air France communicating their plans to increase bonuses for Pieter Elbers and Benjamin Smith, while KLM-Air France is receiving financial support from the government to survive the corona crisis. The purpose of the dataset is to monitor the Twitter opinion on this situation. The dataset was set up in a way that allows for sentiment analysis, a method that could be used to answer research questions regarding this dataset like:

- “What is the sentiment on Twitter concerning KLM-Air France’s intentions to provide bonuses while receiving government support?”
- “Does the type of emotion (angry, happy, sad, etc.) expressed in tweets differ across levels of user activity on Twitter (low, medium, high) regarding the bonus intentions of KLM-Air France while receiving government support?”
- “Do mainly heavy users of Twitter (vs non-heavy users) make posts regarding the intentions of KLM-Air France to provide a bonus to their management while receiving government support?”
- “Does the type of emotion expressed in tweets differ for tweets that are directly aimed at another user with an (@user) than tweets that are a general post not directed at a specific user in regards to the bonus intentions of KLM-Air France while receiving government support?”

1.2 Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Group 20 of the Research in Social Media course provided by the University of Tilburg.

1.3 Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The tools that were required for the creation of this dataset were supplied by the University of Tilburg

1.4 Any other comments?
With performing sentiment analysis it is important to differentiate sarcastic tweets from non-sarcastic tweets. To make this distinction, the most suitable approach is to use Deep Learning structures like CNN or LSTM.

==========
2. COMPOSITION
==========

2.1 What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances in the dataset represent tweets that were posted on Twitter. A tweet can contain emoji's, videos ,images, other tweets and can consist of a maximum of 280 characters.

2.2 How many instances are there in total (of each type, if appropriate)?

The dataset consists of 1903 instances of the same type, tweets.

2.3 Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

While collecting tweets through an API for empirical research it cannot be assumed that the dataset contains all possible instances due to censorship

that may be enforced by twitter. Keeping this in mind the dataset aims to be a sample of all twitter users that discussed the event while it was active. Instances that do not contain the selected or any hashtags were not collected as they are not easily obtainable.

2.4 What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance contains unprocessed text. A typical instance contains a large amount of metadata. This metadata consists of the time the tweet was created, The identification twitter uses for the tweet. The information from the user's profile such as, the location given by the user, the username, the user profile's description, the user's amount of followers, amount of friends, amount of favorites, amount of statuses, the data the account was created, further personalized user information. A typical instance also contains the content of the tweet. As well as further information concerning the tweet, such as its retweet status and the medium that was used to post the tweet.

2.5 Is there a label or target associated with each instance? If so, please provide a description.

Each tweet contains at least one of the following hashtags: "#klm, #topman, #AirFrance, #bonus".

2.6 Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

The instances contain tweets, together with data on the individuals that posted the tweets, including profile data. The data on the content of all the tweets are complete, however users may have adjusted their privacy settings. As such geographical data concerning the user that posted the tweet may be missing.

2.7 Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

In the dataset there exists a relationship between all the twitter users that made use of the same hashtag. Between a smaller number of users a link exists through their tweet being a retweet of another user.

2.8 Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

Data splits are not required, since the event cannot be split up into different sections of content. Due to the nature of the event and the data collected, any validation or testing is possible over different users.

2.9 Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

There are no sources of error in the data set. However, there are some redundancies. There are tweets within the dataset that are not related to the topic we aimed to collect data for. For example, the hashtag ‘bonus’ was used for scraping the data. However, there are some tweets in the dataset that contain this hashtag, but do not talk about the proposed bonuses at KLM - Air France and therefore have no use. A fair amount of tweets that come from this hashtag are compromised as many belong to spam bots that promote casinos and the like, as such these tweets should be curated before the dataset is used. The risk of redundancies also applies to the other hashtags, but are less likely to result in redundancies.

2.10 Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained, it contains images and links that rely on twitter or external websites to respond correctly, however the textual content will always be accessible.
A} There are no guarantees that external websites will keep functioning, it can be assumed that twitter itself will continue functioning.

B} There are official versions of the complete dataset, called twitter Firehose, the full twitter database. However these archives are not easily accessible, as they are being held privately by twitter.
C} There are no external resources in this dataset that might hold any restriction.

2.11 Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

The dataset does not contain data that can be considered as confidential, all data is made public on Twitter by the users. When data on geolocation of users is available, users have given confirmation that this information is made available to the public.

2.12 Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

It can be assumed, as twitter enforces a Sensitive media policy, that the tweets contained in this dataset will not contain excessively gory, violent or adult content.

2.13 Does the dataset relate to people? If not, you may skip the remaining questions in this section.

The dataset does indirectly relate to people, every instance (tweet) is posted by a particular user.

2.14 Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

The dataset does not identify any subpopulations. The instances collected have been posted by individuals with different population characteristics.

2.15 Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Within the dataset it is easily possible to directly identify individuals, since the user-information contains the name users have used to apply to Twitter. However, from inspection of the data it can be noticed that some of the users do not register with their real name.

2.16 Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

The dataset contains publicized information from individual users, as such any personal information they have personally shared can be assumed to be in the dataset. For the personal location the users given city can often be seen.

2.17 Any other comments?

==========
3. COLLECTION PROCESS
==========

3.1 How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data was directly observable from the twitter platform.

3.2 What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

In order to collect the data the Twitter API was used. To acquire actual data from Twitter a script was used using the Python programming language. For the script a library called “Tweepy” was used. Tweepy was used to ease the authorization process and to import the data from twitter to store it in a .json file. Through manual curation the tweets were validated to contain valid information.

3.3 If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The dataset is not a sample from a larger set. We collected instances without the use of a sampling strategy, but collected all tweets over the course of eight hours from all possible users.

3.4 Who was involved in the data collection process (e.g., students, crowd workers, contractors) and how were they compensated (e.g., how much were crowd workers paid)?

The people involved in the collection of the data are all the members of team 20 of the Research in social media course at Tilburg University. All these members are students and they received no compensation for their efforts.

3.5 Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The dataset consists of instances collected from 08:48:35 to 16:19:25 on 23-04-2020. The data was collected between 10:30 and 16:30 on 23-04-2020. Considering this collection timeframe data that was submitted to twitter before the collection started were also gathered. However this seems to be limited to the earliest instance in the dataset at 08:45:35 on the same date the data was gathered.

3.6 Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

During the collection of the dataset no ethical review processes have been conducted. Within the timeframe we collected data, all instances (tweets) have been recorded. It can be expected that the tweets were reviewed according to the documentation available in https://help.twitter.com/en/rules-and-policies/media-policy. However the specific impact of these policies in this dataset are unknown.

3.7 Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.

The data does not directly relate to people. It does indirectly contain the information publicized by people through their account on Twitter.

3.8 Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data was collected through the Twitter platform on which individuals directly published the content of these tweets.

3.9 Were the individuals in question notified about the data collection? If so, please describe(or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
Users of Twitter do not receive a notification detailing that their information was collected as they publicized this information on twitter according to the terms of service that is given in 3.10.

3.10 Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Yes, users of Twitter accepted the terms of service available here: https://twitter.com/en/privacy,to make use of the Twitter platform. Users have full control over the data they share and have the ability to make their account private and inaccessible to third party data collectors.

3.11 If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

The mechanism that Twitter implemented to have users revoke their consent to the collection and use of their data is the deactivation of their account and thereby discontinuing the use of the Twitter services: https://twitter.com/settings/safety

Twitter informs users with the following statement in regards to discontinuing their accounts: “You may end your legal agreement with Twitter at any time by deactivating your accounts and discontinuing your use of the Services. See https://help.twitter.com/en/managing-your-account/how-to-deactivate-twitter-account (and for Periscope, https://help.pscp.tv/customer/portal/articles/2460220) for instructions on how to deactivate your account and the Privacy Policy for more information on what happens to your information.”

3.12 Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

3.13 Any other comments?
No.

==
4. PREPROCESSING/CLEANING/LABELING
==

4.1 Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

4.2 Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

4.3 Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

4.4 Any other comments?

==
5. USES
==

5.1 Has the dataset been used for any tasks already? If so, please provide a description.

No.

5.2 Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

No.

5.3 What (other) tasks could the dataset be used for?

This dataset can be used for network analyses of structural patterns of connectivity and conversation among users. As there are only 1903 tweets in this dataset it would be a small part of the data that is required for network analyses. Besides that the dataset can be used for the reason it was collected in the first, a sentiment analysis in regards to the potential bonuses for the top executives at KLM - Air France.

As the data collected show the amount of tweets and the creation date of the profile the dataset can be used to discern between heavy and non-heavy users of twitter.

5.4 Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

The dataset stereotypes towards active posters on Twitter. It is advised to enrich the dataset with other sources to get a richer, contextualized understanding of the views that people hold concerning KLM-Air France’s decisions.

5.5 Are there tasks for which the dataset should not be used? If so, please provide a description.

The dataset does not give an accurate representation of the opinion of the population on the bonus intention of KLM as the instances are only collected

from people using twitter. Therefore the dataset should not be used to make conclusions about the opinion of the whole population on the bonus intentions of KLM-Air France.

5.6 Any other comments?
No.

==========
6. DISTRIBUTION
==========

6.1 Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which a dataset was created? If so, please provide a description.

6.2 How will the dataset will be distributed(e.g.,tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

6.3 When will the dataset be distributed?

6.4 Will the dataset be distributed under a copyright or other intellectual property(IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU (Terms of Use), as well as any fees associated with these restrictions.

6.5 Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

6.6 Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

6.7 Any other comments?

==========
7. MAINTENANCE
==========

7.1 Who is supporting/hosting/maintaining the dataset?

7.2 How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

7.3 Is there an erratum? If so, please provide a link or other access point.

7.4 Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

7.5 If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

7.6 Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

7.7 If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

7.8 Any other comments?