Research Proposal

Lonneke Langeveld, S3225143

April 2019

1 Introduction and background

For my thesis, I have chosen the topic of 'Stance analysis on Twitter', in line with Mohammad, Kiritchenko, Sobhani, Zhu, and Cherry (2016b). This means that each tweet you collect gets assigned to a 'target' (a person or an entity, such as 'Donald Trump' or 'legalization of abortion'), and the tweet is then classified as being in favour of, against, or neutral towards the target.

Mohammad, Kiritchenko, and Sobhani have done numerous stance analysis tasks (2017; 2016; 2016a), all related to the SemEval 2016-task (Mohammad et al., 2016b). The SemEval-task was actually divided into two tasks; Task A, and Task B. The difference between these tasks was the training data. For Task A, the training and test data were on the same topic. For Task B, no training data was available, but users were allowed to use training data from Task A to train their system.

It is clear that a target is needed to perform stance analysis. I have chosen a target that is current and relevant, which was one of my requirements, but that also *means* something to me. The target I have chosen is the #MeToo-movement, as it is not only both of these elements, but also a controversial topic - which is quite beneficial for stance analysis.

I want to research whether a tweet's author's gender correlates with their opinion - that is, to see if it is easier to predict one's opinion if you know their gender. My research question is "Will including gender as a feature in a classifier influence its ability to classify stance on #MeToo-tweets?", with the following subquestions:

- Are the stances of men and women toward #MeToo siginificantly different from one another?
- What is the difference in the classifier's performance when we do and do not include gender as a feature?
- What is the difference in the classifier's performance when we do and do not use #MeToo-training data (so only #MeToo-test data)?

2 General approach

I am going to do stance analysis on tweets, as had been mentioned a numerous amount of times in the introduction. My target is #MeToo. In the SemEval-task, they used 700-1000 tweets per target. Based on this number, I plan to use a thousand annotated tweets as data. I will collect the data through the Twitter API, pre-process the data, annotate the data, and run an inter-annotator agreement study. After this, I will have a corpus, on which I will run a classifier. After that, I will evaluate the classifier and analyze the data.

Below, I will explain each of these steps in detail.

- **Data collection:** My data will be English tweets. I will find these tweets using Tweepy, a Python wrapper for the Twitter API.[1] I will collect more Tweets than usual, as there are two things that might make a tweet unusable; no gender classification from their username, and its context. I will explain both of these in more detail below. I will make sure that the tweets are the full tweets (max. 280 characters), and not in the shorter format (max. 140 tweets). If I do not do this, tweets that are longer than a 140 characters will be continued on the Twitter website itself, using a URL to link to this, leading to incomplete tweets. Luckily, this is easy to prevent by changing certain search-parameters. In my search query, I will already exclude retweets and replies. Retweets will be excluded because I would not want 'double tweets' to skew the data. I will not use replies, as replies usually require context, and context is not a parameter that I will include in my research. My search query will also include,the fact that the tweets have to be in English. The search query will be '#MeToo movement'. The reason for using the word 'movement' is explained in the 'Tweet content'-section below. I have tested whether this search query will actually yield enough spare tweets; I could collect 5000 tweets with ease. I should be able to gather more than enough (spare) tweets.

- **Pre-processing:** As mentioned above, I will exclude replies and retweets. I will keep all the words after hashtags (so '#MeToo' will become 'MeToo', and the same for any other hashtags a tweet might contain), as hashtags are often used as part of a sentence, and removing them would make the sentence grammatically incoherent. I will keep emoji's, as they might be a good indicator towards one's opinion. I will convert them to something meaningful, for example using Emoji for Python[2]. I will also be removing URLs, as the URLs themselves add no meaning to a tweet.

- **Classyfing gender:** As is clear from my research question, gender is an important part of this research. I will have to classify each tweet's author's gender. To do so, I will use a list of proper names, using this list[3]. I will go through each of the names on this list (manually), to decide whether

[1] http://www.tweepy.org/
[2] https://pypi.org/project/emoji/
[3] https://github.com/smashew/NameDatabases/blob/master/NamesDatabases/first%20names/us.txt

the name is male or female, and so form a separate file for female and male names. If a name can be either (such as 'Robin' or 'Taylor'), it will be excluded.

I will then compare each tweet's author's username to these lists, to see if there is a male or female name in there, and then classify it as that gender. This will happen through a program I will write. This is also where the spare tweets will come in handy: if there is no name in the username, I can exclude the tweet, and grab a new one. This way, gender is classified easily for each tweet, which is necessary for the research.

- **Tweet content:** If I want to classify stance for a tweet, I will have to make sure that each tweet actually *speaks about* the #MeToo-movement, instead of just mentioning it. To do this, I will go through each of the thousand tweets manually (after a gender has been assigned), and see whether the tweet meets this requirement. This is the other part where the spare tweets are used - if a tweet does *not* meet this requirement, I can grab another tweet with ease.

- **Annotation:** The stance annotation will have to happen manually. Gender annotation happens automatically (as shown above). To do these annotations, Excel can easily be used: I put the tweet-ID and the tweet-text in two separate columns, and a third can be filled in with the stance annotation. Gender and username will not be shown during these annotations, as that might (subconsciously) influence the annotator's classification. This could negatively impact the study. I will add gender back to the file later, going by tweet-ID.

  To annotate, we use the guidelines used in Mohammad et al. (2016b), which can be found here.[4] Annotators will classify a tweet as being *in favour of* (FAV), *neutral towards* (NEU), or *against* (AGA) the #MeToo-movement according to these guidelines.

  I ran a small test on a hundred tweets, where I annotated them on my own. I wanted to see whether or not the data would be balanced. I found that the tweets are relatively balanced, with all the categories (FAV/NEU/AGA) containing at least 25 tweets. The thousand tweets should, then, be quite balanced. If this is not the case, however, I can (again) use spare tweets to balance the data.

- **Inter-annotator agreement:** I will 'recruit' someone (one person) to help me with the annotation of the tweets. We will annotate according to the previously mentioned guidelines. We will annotate 300 tweets. If there are any conflicts, these will be resolved. After this, I will continue annotating the data independently. To see what the inter-annotator agreement is, Cohen's kappa will be used.

[4] http://alt.qcri.org/semeval2016/task6/data/uploads/stance-question.pdf

- **Corpus:** The final corpus will consist of a thousand tweets. Each tweet will have a tweet-ID, its stance (FAV/NEU/AGA), and the author's gender.
- **Classifier:** I will most likely use the `scikit-learn`-SVM classifier. It was the baseline in Mohammad et al. (2016b), which was established in Mohammad et al. (2017), and it seems that, as it stands, it is an excellent classifier. The baseline for my research will be decided by popular vote - how big is the chance that when you just guess the majority group (in favour, neutral, or against), you are right?

  I will do both type A- and type B-types of tasks, as they did in Mohammad et al. (2016b). With type A-tasks, I will use, as they did, 70% (so 700 tweets) training data, and the rest of the tweets (the other 300) being test data. With type B-tasks, all #MeToo-data will be test data. For training data, we can, for example, use the data from Mohammad et al. (2016b), and then especially the 'feminism'-part and/or data from the Automatic Misogyny Identification-task[5].
- **Types of classifier:** I will do both type A- and type B-tasks. These tasks have been described in the introduction.
  - **Type A:** For the type A-task, I will use 70 percent of the data as training data, and the other 30 percent as test data. This is what they did in Mohammad et al. (2016b).
  - **Type B:** For the type B-task, all the #MeToo-data will be test data. Training data can, for example, be taken from the SemEval-task, and then especially the 'feminism'-part. The #MeToo-movement has been defined as *"one of the most high-profile examples of digital feminism we have encountered"* (Mendes, Ringrose, & Keller, 2018). We can therefore subtly assume that tweets that are positive towards feminism will be in favour of the #MeToo-movement as well. Another possible data source for the Type B-training data could be the data from the Automatic Misogyny Identification-task.[6] This task tagged tweets as being misogynist or not. It is common knowledge that misogynists are not feminist, and will certainly not support the #MeToo-movement. We could then use tweets tagged as 'not misogynist' as, generally, being in favour of the #MeToo movement (and vice versa).

3 Expected output and evaluation

The output of my research will be a classifier, as well as a corpus. The classifier will classify #MeToo-tweets as being in favour of, neutral toward, or against the #MeToo-movement. I will classify with and without gender, to see whether this has a significant effect on the classifier's scores. To see how well the classifier

[5] https://amievalita2018.wordpress.com/data/
[6] https://amievalita2018.wordpress.com/data/

does its job, I will use the same evaluation measures as in Mohammad et al. (2016b). The measure is as follows:

$$F_{avg} = \frac{F_{favour} + F_{against}}{2}$$

where F_{favour} and $F_{against}$ are calculated as shown below (with P being precision and R being recall):

$$F_{favour} = \frac{2P_{favour}R_{favour}}{P_{favour} + R_{favour}}$$

$$F_{against} = \frac{2P_{against}R_{against}}{P_{against} + R_{against}}$$

References

Mendes, K., Ringrose, J., & Keller, J. (2018). # metoo and the promise and pitfalls of challenging rape culture through digital feminist activism. *European Journal of Women's Studies*, *25*(2), 236–246.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016b). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)* (pp. 31–41).

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X.-D., & Cherry, C. (2016a). A dataset for detecting stance in tweets. In *Lrec.*

Mohammad, S., Sobhani, P., & Kiritchenko, S. (2017). Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, *17*(3), 26.

Sobhani, P., Mohammad, S., & Kiritchenko, S. (2016). Detecting stance in tweets and analyzing its interaction with sentiment. In *Proceedings of the fifth joint conference on lexical and computational semantics* (pp. 159–169).