Hate Speech Detector

Hate speech tweets classifier with application of neural networks

Karol Kowalski (220985), Jakub Kałużny (218159)

Definition

Hate speech denotes verbal or nonverbal attack against a person or a social group. This concept does not concern only the **insults on the background of protected attributes** such as race, religion or national origin. It concerns every **abusive, humiliative or bullying language**.

Dataset and Classifier

English training data were obtained from **Davidson et al.** Github webpage (https://github.com/t-davidson/hate-speech-and-offensive-language) and polish from **Poleval 2019** website (http://2019.poleval.pl/index.php/tasks/task6). English dataset contains 77.43% hate and 22.57% non-hate speech examples. Polish dataset contains respectively 8.92% and 91.08%.

The test data were scrapped from Twitter. Appropriate tweets were searched by hashtag 'Iran' as english data and 'tylkoniemownikomu' as polish. Then it has been performed following analyses. First, which hashtags clearly denote hate-speech classified tweets. Second, what are the percentages of hate- and non-hate-speech tweets denotation for 10 most frequent ambiguous hashtags.

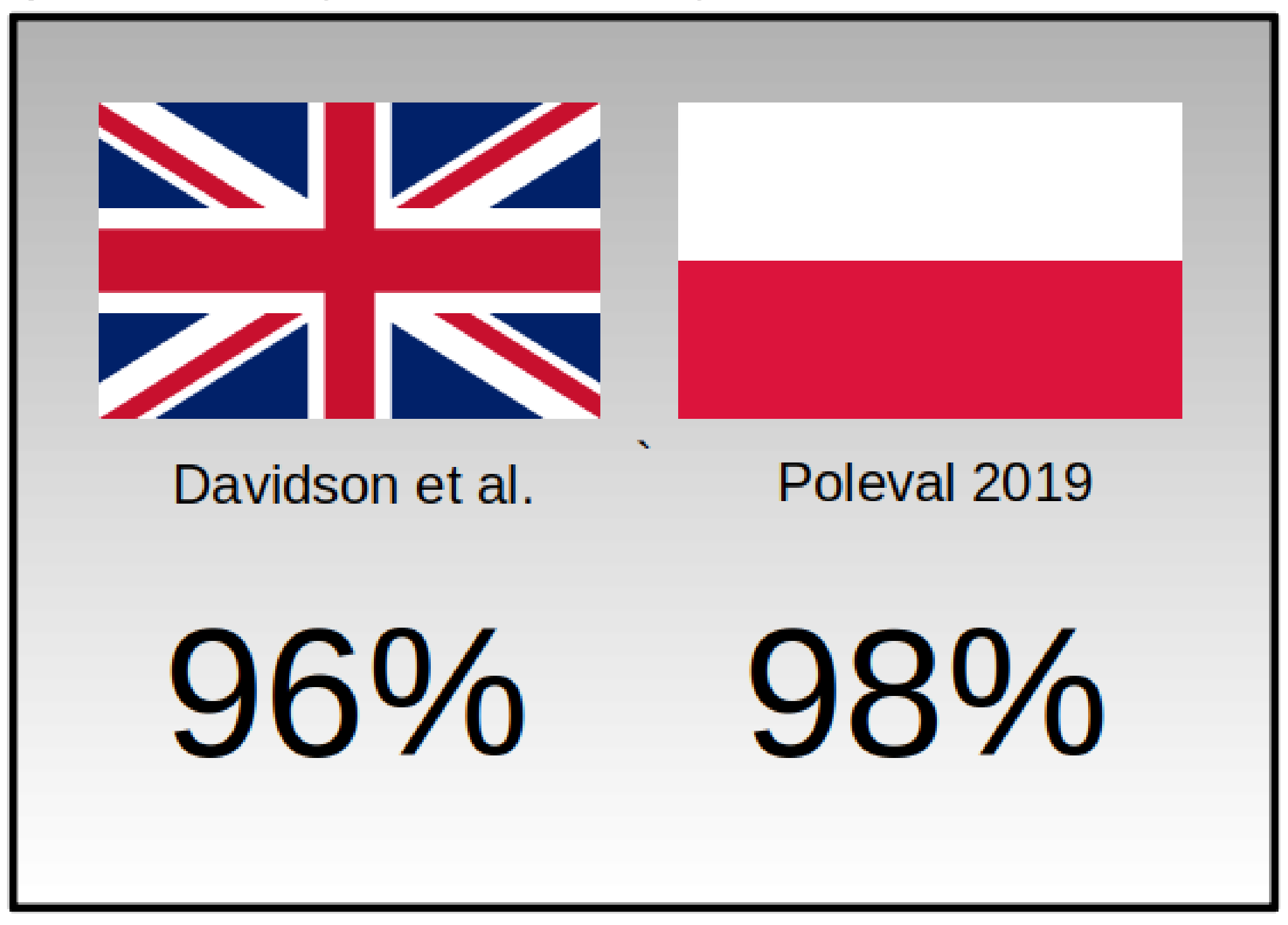

Figure 2: *LSTM neural network classifier accuracies for Davidson et al. and Poleval 2019 tweets.*

Phases of Project

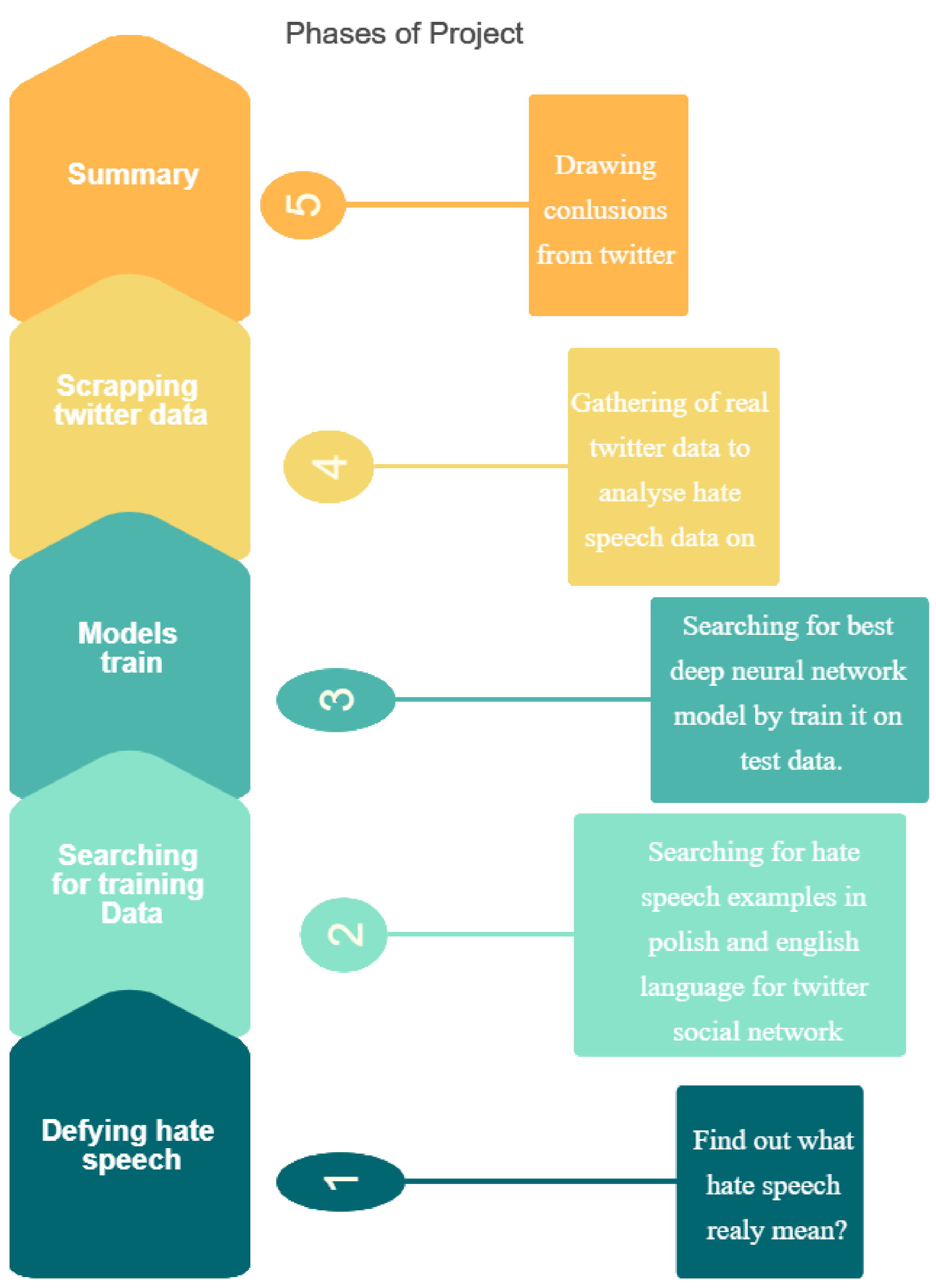
Phases of Project
Summary
5
Drawing conlusions from twitter
Scrapping twitter data
4
Gathering of real twitter data to analyse hate speech data on
Models train
3
Searching for best deep neural network model by train it on test data.
Searching for training Data
2
Searching for hate speech examples in polish and english language for twitter social network
Defying hate speech
1
Find out what hate speech realy mean?

Figure 4: *Main project phases*

Hashtags clouds

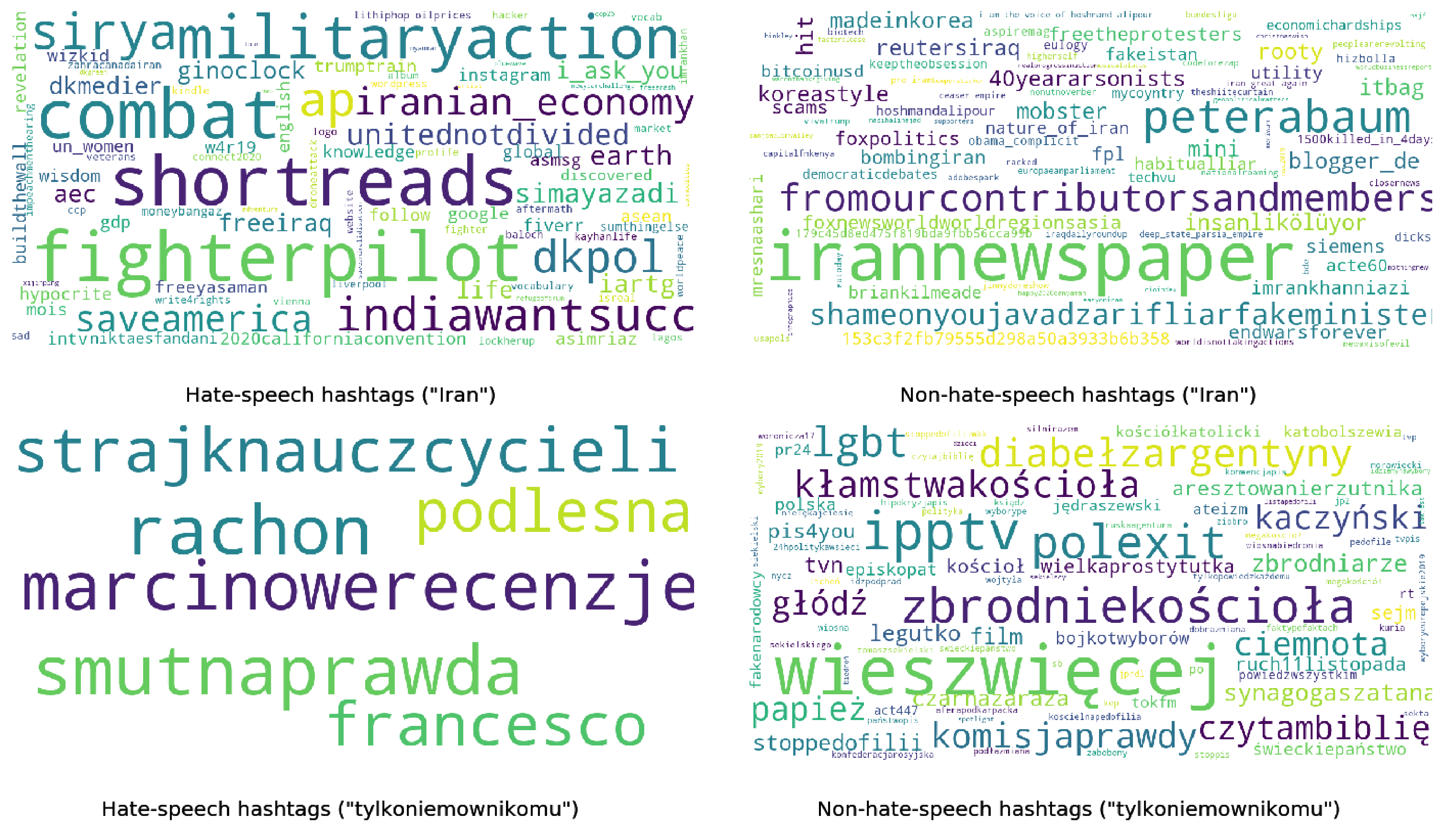

Figure 1: *Examples of hashtags which appeared in hate- and non-hate-speech for english and polish tweets.*

Thanks to these clouds it could be found out which hashtags are strongly corelated with hate or non-hate speech. Judging by above analyses the **polish users of twitter are less prone to use hate speech than english users.** Whereas considering the hashtags contained in non-hate-speach classified polish tweets the conclusion is that the **model might have been tested on irrepresentative data** because some of the hashtags rather denote hate-speech tweets.

Hashtags analysis

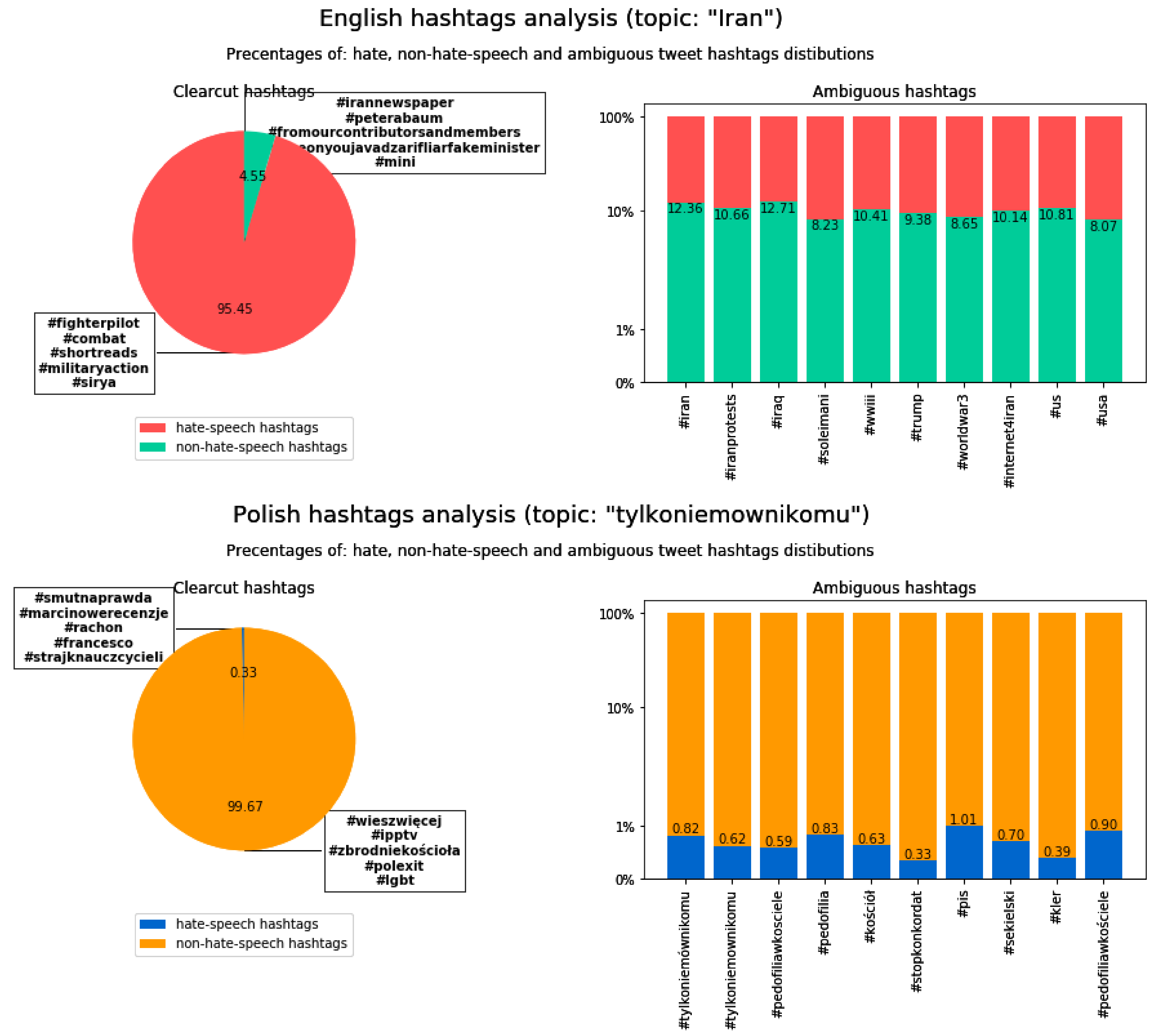

Figure 3: *Ambiguous and clearcut (hate- or non-hate-speech) hashtags analysis*

The above hashtag amount analysis shows that (just like by wordclouds) **english users are more prone to use hate speech in tweet than polish users**. These results however might also be **biased with imbalanced or irrepresentative tweet data**.