Discourse Analysis of Events on Twitter

Degree programme: BSc in Computer Science | Specialisation: Data Engineering
Thesis advisor: Prof. Dr. Erik Graf
Expert: Daniel Voisard

As the use of social media, such as Twitter, rises every year, the relevance of the discussions taking place on these platforms also increases for journalists or other professionals analysing the discourse of events. However, analysing millions of tweets is a very time consuming task, which also requires a lot of resources. So, this is where machine learning can provide an optimal solution.

Céline Hüttenmoser

Introduction

The focus of this thesis is to analysis the discourse of events on Twitter using Natural Language Processing. Considering that, several techniques have been compared in order to get satisfying results, including a discussion about the challenges such a task can represent.

Methods

Data Collection

During this thesis, the focus has been placed on the Brexit event, which occurred between June 2016 and December 2021. Therefore, the tweets, retrieved through the official Twitter API, were filtered by the word Brexit, either in the hashtags or the text. Giving a clear definition of an event, as well as using a good filtering method of the tweets have been shown to be challenging, but important tasks.

Semi-Supervised Learning

As labelling a large number of tweets is very time consuming and requires a large amount of resources, choosing a semi-supervised method can be helpful. First, the tweets have been labelled according to their hashtags or the keywords. Using this training data, a model using the logistic regression algorithm has been trained. As a second step, a self-training method was used, which allows the training with labelled and unlabelled data. This technique is especially useful when the models need to be updated, with new datasets of tweets, so as to increase the learned vocabulary of the models.

Transfer Learning

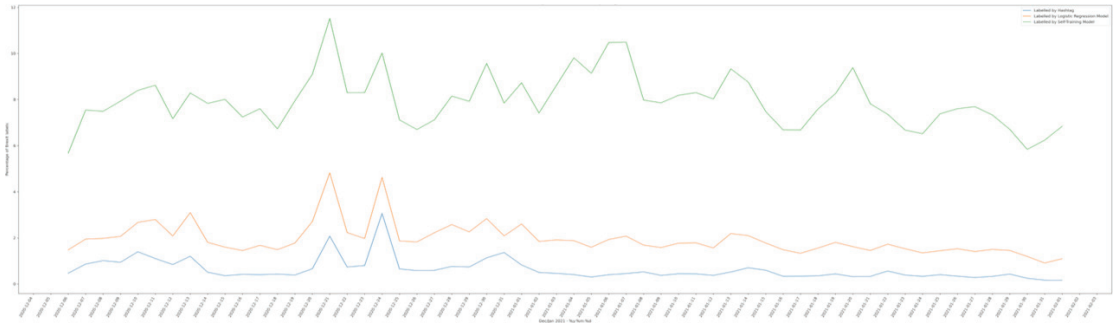
Due to the short amount of text present in tweets, not much context is given, which represents a further challenge for the labelling. Therefore, a pre-trained Glove model and a model, trained with an own dataset using Doc2Vec, have been transferred to a logistic regression algorithm for the training.

Pooled Evaluation

For a qualitative evaluation of the dataset and the models, pooling has been used to retrieve some data which have been manually analysed and labelled by different annotators.

Results

The analysis of the results, and the high accuracy of the models, have shown the feasibility, however, also the challenges of such a task. The three different labelling methods yield different results in terms of proportions. However, the general trend is the same.



Daily frequency of tweets related to the event Brexit between 06.12.2020 – 01.02.2021