Stance Detection in Tweets

W266 Natural Language Processing - Final Project

Alex Dessouky & Tim Spittle

December 7, 2019

Abstract

TBD - Abstract goes here

1 Introduction

1.1 Background

Stance classification is a natural language processing (NLP) task that seeks to automatically determine the "stance" of a statement in relation to a "target". The stance the author of that statement can take is either in **favor** of the target, **against** the target, or **neutral** toward the target. A target can be anything, from a proposition to an idea, a person, a political issue, etc. The stance of any statement can be determined against any target. Since the statement may not address the target at all, it is possible that the author's stance towards the target of evaluation cannot be determined. In certain circumatances where the at-issue target is not addressed directly, the stance of a stamenet can be determined by inference from association with another target.

Given an abundance of online discourse on social media, automatic stance detection of social media content is an appealing task with widespread potential applications. In particular, Twitter is a social media platform with an incredibly fast-paced environment where many tweets are shared back-and-forth addressing a variety of topics and issues. Stance classification of tweets has applications across many domains: marketing industry efforts to measure public opinions on products, political campaign attempts to measure public views on candidates' policies, and efforts by Twitter to identify bad actors (i.e., "trolls"). While there has been significant research in stance classification with respect to debates and online forums, Twitter poses a new challenge as many "tweeters" represent their stance towards a target implicitly and often use figurative language, shorthand, acronyms, or hashtags.

1.2 Objective

The objective of this paper is to explore new methodologies for detecting stance of tweets and attempt to design a classification model that outperforms existing models.

2 Data, Task, & Evaluation

We relied on an existing, pre-labeled Twitter dataset that has been analyzed by competitor models against which we can compare our results. *(Mohammad et al. (2016a))* This dataset was originally compiled for the International Workshop on Semantic Evaluation 2016 (SemEval-2016) where it was used in an evaluation exercise, Task 6. *(Mohammad et al. (2016b))*

The dataset consists of approximately 5,000 tweet-target pairs annotated for both stance and sentiment. The targets may or may not be referred to in the tweets. While sentiment is shown to be beneficial for stance classification, it was not available to competitors at the SemEval-2016 conference. Given this in combination

with the fact that pre-labeled sentiment would not be available in a production environment, the sentiment labels were ignored for our classification model (*Sobhani, Mohammad, and Kiritchenko (2016)*).

The process for collecting and annotating these tweets is detailed in *Mohammad et al. (2016a)* . The authors queried Twitter for topic-specific hashtags in an effort to collect a full corpus of tweets related to the intended targets. These targets include:

Table 1: Target Categoris with Example Hastags

Target	Favor	Against
Atheism	*#NoMoreReligions*	*#GodsWill*
Climate Change is a Real Concern	-	*#GlobalWarmingHoax*
Feminist Movement	*#INeedFeminismBecause*	*#FeminismIsAwful*
Hillary Clinton	*#GoHillary*	*#WhyIAmNotVotingForHillary*
Legalization of Abortion	*#ProChoice*	*#PrayToEndAbortion*

After extracting the results, the authors retained only tweets with the query hashtag at the end of the tweet. The query hashtag was then removed in order to exclude obvious clues for the classification task. Note that while this is a necessary step for the NLP task, removal of the query hashtag may actually change the stance of the tweet (e.g. if the hashtag was a negation of the preceding text).

Tweets were then annotated for stance using CrowdFlower (http://www.CrowdFlower.com)[http://www.crowdflower.com], a website for crowd-sourced data annotation, where each tweet was annotated by at least eight reviewers. The raw inter-annotater agreement was 73.1%; however a cutoff was applied where only tweets with greater than 60% agreement were retained, increasing the average agreement to 81.85%. Lastly, the resulting tweets were split into train and test sets chronologically. (See the **Limitations** section for more on how this process may affect the classifcation task).

Models were trained independently to predict stance for each target separately. Performance was evaluated for each target using a modified F1 score, which took the macro-average of the "favor" and "against" class F1 scores only. This method was chosen to treat the "none" stance as a class that is not of interest (or negative class), though it still will affect the macro-average F1 scores. *[NEED Information Retrieval Source]* For evaluation of model performance across all targets, the micro-average of F1 scores was evaluated (referred to as **F-microT**). This was the score used as the official competition metric. Alternatively, the macro-average of the F1 scores can also be calculated (referred to as **F-macroT**) by averaging the modified F1 scores acheived for each target. *(Mohammad et al. (2016b))*

3 Related Work

3.1 Baseline Models

The baseline model for the SemEval-2016 Task 6 competition was a linear-kernel **Support Vector Machine (SVM)** classifier with three different specifications and features sets used *(Mohammad et al. (2016b))*. The *unigrams* model created five SVM classifiers (one per target) trained using unigram features, the *N-grams* model created five SVM classifiers (one per target) trained using word n-grams of 1-, 2-, and 3-gram length and character n-grams of 2-, 3-, 4-, and 5-gram length, and the *N-grams-combined* model created one SVM classifier trained on the combined data for all targets, using the say n-gram features as described above.

SVM was chosen as a baseline since it is effective on text categorization tasks and robust on large feature spaces. *Sobhani, Mohammad, and Kiritchenko (2016)* expanded on the *N-grams* model to also include word embeddings and sentiment lexicon features. Note that the word embeddings in this model were trained using the "full corpus" of nearly 2 million tweets extracted by the authors that comprised the original dataset using query hashtags. These additional tweets were not provided with the data, therefore the strength of

this model may be artificially inflated through the context these word embeddings were able to capture by virtue of being trained with exclusive tweets that align precisely with the task at hand.

3.2 Sem-Eval 2016 Competitors

Several of the competing teams within the SemEval-2016 Task 6 competition took a **neural network** approach to the challenge. *Zarrella and Marsh (2016),* the winner of the competition, leveraged transfer learning to perform the classification task. The team sampled 200 million+ relevant tweets to create an auxiliary 'hashtag prediction task', which was used as a projection into a Long Short Term Memory (LSTM) - Recurrent Neural Net (RNN) layer. *Wei et al. (2016)* and *Vijayaraghavan et al. (2016)* both implemented convolutional neural networks, however, these models did not achieve the same level of success as *Zarella et. al.*

There were also a handful of teams that took a **featured-based** modelling approach. *Tutek et al. (2016)* used 'lexical features' such as n-grams, average word length, and number of hashtags as inputs to an ensemble classifier rooted in logistic regression, gradient boosting machines, random forests, and support vector machines. Similarly, *Bøhler et al. (2016)* used a voting classifier that took input from two multinomial Naive Bayes classifiers, one trained on word bi-grams while the other was trained on character tri-grams, and a logistic regression classifier trained on GloVe vectors. *Bohler et. al* also experimented with various additional features such as the presence or absence of negation and length of tweets.

3.3 Subsequent SemEval Progress

Since Sem-Eval 2016, there have been two additional contributions to this task of note. Both *Sun et al. (2018)* and *Du et al. (2017)* leveraged **neural attention networks** to perform stance classification, with each team achieving higher final F1 scores than the winner of the initial competition. Although several of the competing teams found effective models, none were able to achieve higher final F1 scores than the SVM baseline from *Mohammad et al. (2016b)* . This also holds for *Sun* and *Du*'s work after the conference ended.

3.4 General State-of-the-Art NLP Methodologies

Mayfield and Black (2019) applied *Devlin et. al*'s BERT contextualized word embeddings to a stance classification task attempting to predict whether a Wikipedia user preferred to 'Keep' or 'Delete' a specific post based on their comments. *Mayfield et. al*'s results show that BERT out-performed their baseline classifiers, which used GloVe and Bag-of-Words embeddings as inputs.

Further, *Ma (2019)* experimented with various infrastructures built on top of BERT outputs for a binary classification task using Twitter data. The overall goal of *Ma*'s experiment was to determine whether a given tweet was 'on-topic' or 'off-topic' in the context of disaster management. *Ma* found that a bi-directional LSTM taking in the full BERT sequential output outperformed the standard BERT pooled output used for classification tasks.

4 Model

Next, we describe the model we implemented to approach SemEval-2016 Task 6. After reviewing the work performed by contributing members of the competition, we hypothesized that a neural network which effectively captures the context of the limited characters and words contained in a tweet would perform well in classifying stance from Twitter data. Words contained in tweets cannot always be taken at face value. Hashtags, sarcasm, irony, and other forms of rhetoric are frequently found in tweets, and we would want our model to effectively capture these nuances since it contains valuable information about the author's stance.