TBD, Twitter Sentiment Analysis

Andrea Canonica, Oliver De La Cruz, Claudio Röthlisberger, Simon Scherrer

June 18, 2017

Contents

1 Abstract

NOTE: the abstract will be written after we have finished the other sections, as suggested in the instructions for writing a scientific paper provided on the lecture website

2 Introduction

- sentiment analysis is a current research topic
- although a specific problem, wide range of applications (publishing, economy, artificial intelligence, etc.)
- in this particular case we analyzed twitter posts, classifying them as either positive or negative

3 Our Method

3.1 Twitter Dataset

- the dataset was provided on the course website
- 2 million posts, with approximately 1 millions posts per class (positive and negative)
- other datasets are available (e.g. from Sentiment140) *NOTE: our dataset is probably a subset of this*
- emoticons were removed
- usernames (i.e. @-mentions) were replaced with a token
- urls were also replaced with a token
- hashtags were not altered
- similar to the procedure described in sections 2.2 and 2.3 of the paper Twitter Sentiment Classification using Distant Supervision

3.2 Preprocessing

3.2.1 Challenges

- The language used in tweets differs from standard english. To fit the message into the 140 character prescribed by the system, an author may use a style similar to telegrams and abbreviate words. These abbreviations can be rather unconventional and only have a meaning in the context of the message.

- Because of the informal nature of communication, grammar and orthography are not valued as much as with other means of communication.
- Tweets may contain hashtags which convey a meaning, but are not part of any sentence (e.g. #blessed, or #thatsmylife).

The paper The Role of Pre-Processing in Twitter Sentiment Analysis tries to address some of these challenges.

3.2.2 Hashtags

- *NOTE: Andrea's code treats hashtags as a distinct word of the corpus*

3.3 Word Embedding

3.4 Classification

- convolutional neural network (CNN)

4 Results

4.1 Preprocessing

NOTE: show the effect of the preprocessing on the quality of the classifcation

4.2 Classification

4.2.1 Baseline

NOTE: what will be use as the baseline to compare our method to?

5 Discussion

- strenghts and weaknesses
- implications on the application

6 Summary

NOTE: show our contribution