**Master Thesis Artemiy**

**Research Question:**

Main question:

How to effectively use news articles in evaluation of the company’s current financial and/or public perception state?

Subsequent questions:

What are the current state-of-the-art methods of sentiment analysis for unlabeled data?

How do pretrained word/sent/doc embeddings and other shallow algorithms for features extraction differ in terms of performance on news articles dataset?

How does each of them differ in comparison with performance on benchmark datasets (benchmarks can be obtained here: <http://www.aclweb.org/anthology/W17-5202>)

**Expected outcome:**

Implementation part: an ML model with the capability to perform sentiment analysis on a set of documents (news articles) with a corresponding pipeline for data cleaning and consumable results visualization. The code should be able to give the user a basic statistics on the company’s performance over different news sources or/and the timeline of publish.

Evaluation part: overview of different word/sent/doc embeddings and their performance on the given dataset. Recommendations on further research and limitations of generalization capabilities of the model are going to be presented as well.

**Approach:**

1. Aggregate and convert the dataset into suitable format
2. Clean the dataset from the certain datapoints of minimal interest for the research based on criteria;
   1. Article is too short
   2. Article is too long
   3. Article is irrelevant and/or talks about the company briefly
   4. Article is about different company
3. Present descriptive statistics about the dataset in order to better understand the structure and possible ways for improvement for it
4. Choose the 3 state-of-the-art best performing word/sent/doc embeddings for our investigation based on the previous benchmark researches
5. Perform analysis on our dataset
6. Perform benchmarking comparing how the models generally perform on our dataset and how the results differ from the ones found in other benchmarking papers

**Timeframe:**

Start: 15.06

Hand-in: 15.08

**Table of Contents:**

|  |  |
| --- | --- |
|  | *# of pages* |
| 0 Abstract | 1 |
| 1 Introduction | 7 |
| 2 Literature Review | 12 |
| 3 Hypothesis Development | 7 |
| 4 Method & Data | 12 |
| 5 Analysis | 10 |
| 6 Discussion  *Implications for theory and management*  *Limitations and Outlook* | 7 |
| 7 Conclusion | 2 |
| 8 References | 3 |
| 9 Appendix *(if required, e.g. applied R/Python code, background information, etc.)* | 15 |