Similarity measurement of news articles and party manifestos.

[Proposal] *

Florian Niederhuber
Universität Passau
niederhu@fim.uni-passau.de

Jürgen Rottmeier
Universität Passau
rottmeie@fim.uni-passau.de

Felix Steghofer
Universität Passau
steghofe@fim.uni-passau.de

Keywords
natural language processing, text similarities, party manifestos, news articles

1. INTRODUCTION

Media has always been influenced by political movements and therefor by political parties. An author may also sympathize with a specific political party and shares more or less opinions that party represents. Parties do usually publish so called party manifestos where they want to announce their current position to the social life and present their solutions to existing problems. These manifestos are published e.g. for a government period which is often four or five years. In this work we want to classify a news article to a political position using different algorithms and concepts which find similarities between a news article and party manifestos. With these results we try to assigning an author to a political party by analysing a set of articles that author has published. It is also interesting if the same can be applied to whole newspapers. For example if a newspaper often releases texts with positions that refer to a political party's manifesto. Figure 1 shows how manifestos can be classified with terms used regarding a topic.

2. CONCEPTUAL DESIGN

The following chapter describes the general approach to achieve the gains introduced before. The approach can be categorized into different steps that will be explained in the following sections and in figure 2:

2.1 Gather Information a Normalization Process

Based on Party Manifestos proposed by each party in the Internet, a parser is created that extracts all words to a raw text file that can be used in further steps to work with. Based on Niesink the following steps are applied to normalize the raw text: [2]

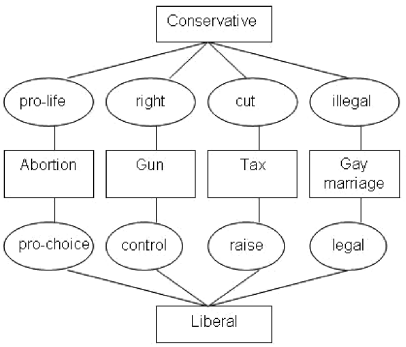

Figure 1: Conservative vs. liberal, [3, Page 33]

- Tokenization
- Lowercase Text
- Punctuation removal
- Stop word removal
- Stemming

2.2 Bag of words and TF-IDF

The extracted tokens from the previous section are used to create a vector for each document whereby the terms are representing the dimensions and the values the weight. In order to get more meaningful weights TD-IDF is used instead of the plain frequency of words.

2.3 Calculating similarity with Pearson Correlation Coefficient

Based on the evaluation of A. Huang the Pearson correlation coefficient is used to determine the similarity between each document[1]. Given the results we can assign news articles to the different parties.

3. OPTIONAL GOALS

- Implement another similarity measurement algorithm. (e.g. Jaccard Coefficient)
- Graphical representation of the results.
- Web platform to process user defined texts and manifestos.

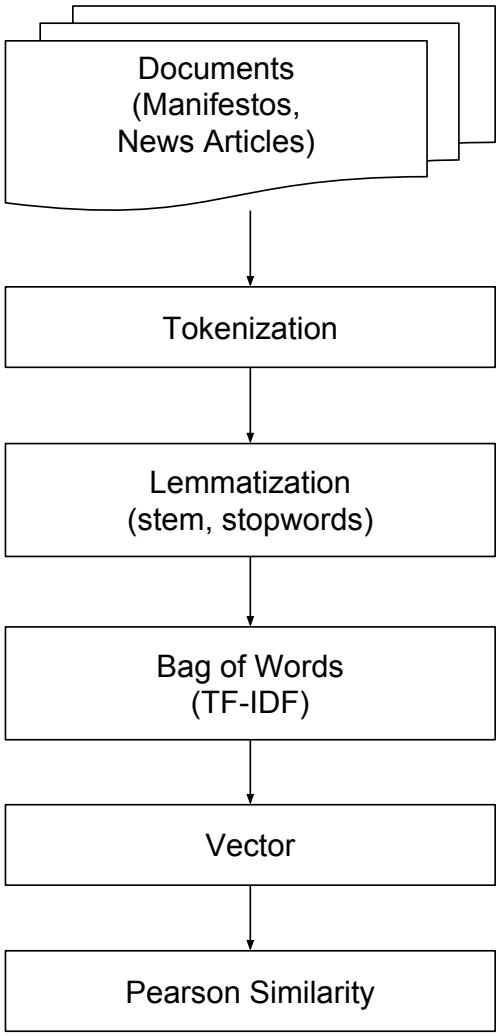

Figure 2: Document processing workflow.

4. REFERENCES

[1] A. Huang. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, pages 49–56, 2008.

[2] S. Niesink. Finding sources of online news articles using textual similarity. *22nd Twente Student Conference on IT*, 22(January 23), 2015.

[3] B. Yu, S. Kaufmann, and D. Diermeier. Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, 5(1):33–48, 2008. `http://beiyu.syr.edu/JITP2008.pdf`.