Your Title

submitted in partial fulfillment for the degree of
master of science

Axel Hirschel
10656146

master information studies
data science
faculty of science
university of Amsterdam

Your Date of defence in the format YYYY-MM-DD

	Internal Supervisor	**External Supervisor**
Title, Name	Dr Maarten Marx	Tom Kunzler
Affiliation	UvA, FNWI, IvI	Open State Foudation
Email	maartenmarx@uva.nl	tom@openstate.eu

Universiteit van Amsterdam

Todo list

Contents

Abstract

Thesis requirements

- Your thesis is written in ACM style with two columns (`documentclass[sigconf]acmart`).
- It is maximally 10 pages long, excluding the title page and the appendix, but including references, figures, etc

1 Introduction

Text classification is considered as one of the most important challenges within natural language processing. Classifying documents is vital, as it enables users to easily query and retrieve useful information. Moreover, it allows the automization of many processes such as spam classification and sentiment analysis. Given the wide variety of application, many algorithms have been developed to tackle the problem (Aggarwal & Zhai, 2012).
Bayesian Classifiers are considered a class of classification algorithms, and they classify document based on word occurences within documents. This word presence is used to calculate the probality that certain documents are part of a topic. The two prominent versions of bayesian classifiers are multi-variate Bernoulli models and multinomial models. Another widely used class of text classificatiers are support vector machines (SVM). Within SVM the algorithm creates linear hyperplanes which split the data into classes based on a bag-of-words representation of texts. Using kernel tricks hyperplanes can be constructed which can find more compex relations than linear (Aggarwal & Zhai, 2012).
Recently, deep neural networks have been employed on classification problems as well. Most notably convolutional neural nets (CNN) have outperformed other methods on baseline classification problems. CNN have been generaly been used on image data, but research in word and document embedding spaces such as Word2Vec (Mikolov, Chen, Corrado, & Dean, 2013) and Paragraph2Vec (Le & Mikolov, 2014) allow the use of CNN on text as well. Transforming words within documents into a multi-dimensional vectors allows the use of convolutional filters, which shift over the documents and detect patterns within documents (Kim, 2014).
In contrast to many of the baseline challenges within text classification, real-world application of classification often involves other challenges as well. Within this research documents of Dutch municipalities are classified, which is a difficult task due to three properties. Firstly, no labelled training data is available, which means that training needs to be done on data from the central document. Secondly, many of the documents are multi-topic and it is interesting to discuss how well algorithms deal with this. Thirdly, within classes a large variety of documents exist, as the documents differ in length and style. The research question and subquestions are thus:

- How well does CNN perform on classifying Dutch municipality documents compared to SVM and Bayesian Classifiers?
  - How does the detail of topics influence the performances of all algorithms?
  - What thresholds are optimal in detecting topics of multi-labeled documents.
  - How well do the performances of the algorithms on the dataset of the central government generalize to the dataset of the municipalities?
  - How can all algorithms be optimized to deal with the large in-class variety of the topics?
  - What are the characteristics of misclassified documents?

Within the next chapter, the literature review, current approaches to the mentioned challenges and the general idea behind the algorithms is further explained. Then the specific set-up for this research is discussed within the methodology, which also provides information on how the research questions are answered. The results of this research is described next, and provides a detailed overview of the performance of all algorithms with various evaluation metrics. Lastly, the answers to the research questions are formulated and the conclusions from this article are discussed.

2 Related Work

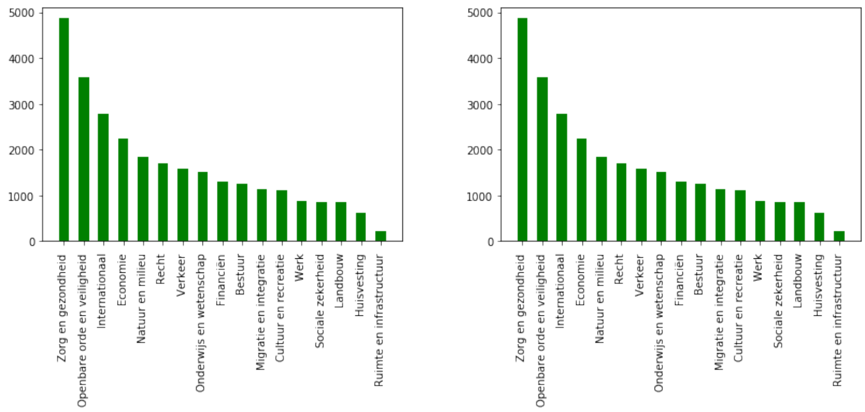
2.1 Document representation

Documents can be represented in multiple ways when used within text classification. One of these methods is bag-of-words, which means that each document is represented as a vector with the length of the vocabulary. Each entry in that vector represents how many times the corresponding word occurs within that document. Often, this representation is expanded upon with TF-IDF, which includes the inverse document frequency of words as well. Therefore the occurences of words within are scaled based on the total amount of documents in which that word occurs.
While bag-of-words has been the standard representation for many years, it does have significant disadvantages. This is foremost caused by the loss of word order within that representation. Also, words with similar meaning, such as ”hate” and ”loathe” are equally far apart within this representation as words with totally different meaning such as ”hate” and ”love”. This means that this representation is less well equiped to deal with nuanced differences in meaning and context.

3 Methodology

3.1 Description of the data

The data used within this project consist of two types of data, both from the government of the Netherlands. The first set consists of over 20,000 questions aksed within the Dutch national parliament each annotated with two labels. One of the 17 broad labels, such as healthcare or education, and one of the 118 more detailed labels, such as elderly healthcare or primary education. Figure 1a and 1b show the distribution of topics within the dataset. These questions also vary in length, as can be seen in Figure **??**. The questions are collected from www.zoek.officielebekendmakingen.nl with a scraper and the set consist of all question asked in 2016 and 2017.



(a) Distribution of the 17 broad topics (b) Distribution of the 118 specific topics

Figure 1: Distribution of topics within the train data

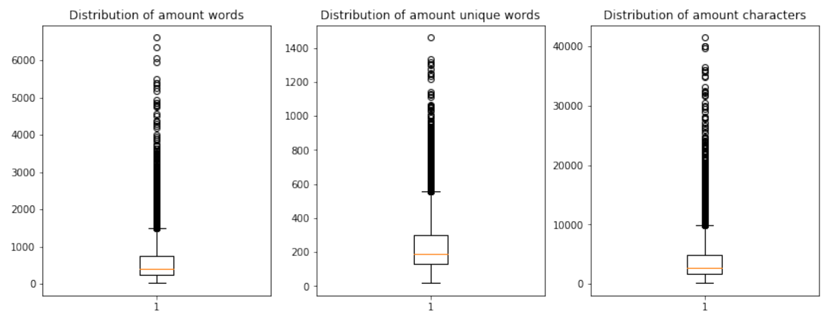


Figure 2: Box plots of the amount of words, unique words and characters within the train data.

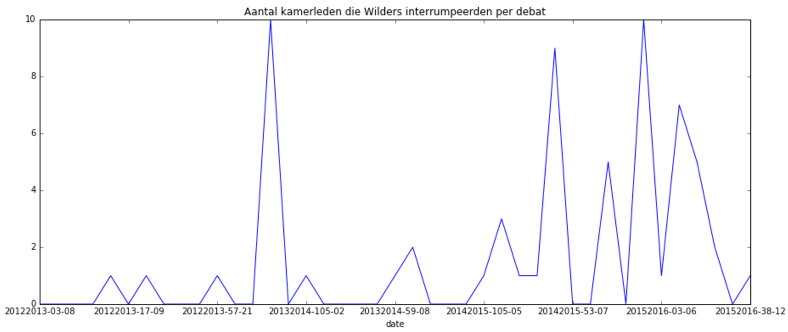


Figure 3: Aantal interrupties van Wilders in de Tweede Kamer door de tijd (periode 2012-2016).

3.2 Wat plotjes en tabelletjes

Zie het IPython Notebook `PandasAndLatex.ipynb` voor de code om vanuit pandas een poltje op te slaan en een dataframe als tabel op te slaan. Het werkt ideaal!

De interrupties van Wilders staan beschreven in Figure 3 en Tabel 1.

date	indegree	interruptie_volgorde
20122013-03-08	0.0	
20122013-07-16	0.0	
20122013-100-03	0.0	
20122013-100-06	0.0	
20122013-17-06	1.0	Pechtold-3
20122013-17-09	0.0	
20122013-21-04	1.0	Pechtold-3
20122013-22-08	0.0	
20122013-32-06	0.0	
20122013-48-23	0.0	
20122013-57-21	1.0	Pechtold-6
20122013-76-03	0.0	
20122013-76-06	0.0	
20132014-05-02	10.0	Roemer-4 Van Haersma Buma-4 Pechtold-4 Slob-5 ...
20132014-06-04	0.0	
20132014-105-02	1.0	Pechtold-10
20132014-105-06	0.0	
20132014-14-03	0.0	
20132014-14-06	0.0	
20132014-52-18	0.0	
20132014-59-08	1.0	Klaver-3
20142015-02-08	2.0	Pechtold-6 Slob-4
20142015-03-06	0.0	
20142015-09-09	0.0	
20142015-100-05	0.0	
20142015-105-05	1.0	Pechtold-2
20142015-111-04	3.0	Pechtold-6 Kuzu-8 Klaver-3
20142015-111-07	1.0	Pechtold-2
20142015-39-71	1.0	Pechtold-2
20142015-41-07	9.0	Samsom-2 Pechtold-3 Kuzu-6 Zijlstra-5 Van Ojik...
20142015-53-07	0.0	
20142015-61-23	0.0	
20142015-79-07	5.0	Klaver-10 Gesthuizen-3 Voordewind-2 Pechtold-6...
20142015-95-06	0.0	
20152016-02-07	10.0	Pechtold-5 Slob-7 Klaver-11 Kuzu-24 Öztürk-1 S...
20152016-03-06	1.0	Pechtold-5
20152016-14-02	7.0	Klaver-9 Roemer-4 Samsom-2 Van Haersma Buma-5 ...
20152016-14-05	5.0	Van Haersma Buma-13 Pechtold-4 Zijlstra-1 Klav...
20152016-27-03	2.0	Segers-4 Kuzu-10
20152016-38-10	0.0	
20152016-38-12	1.0	Klein-2

Table 1: Door wie werd Wilders onderbroken en hoe vaak per debat.

3.3 Methods

Hoe je je vraag gaat beantwoorden.

Dit is de langste sectie van je scriptie.

Als iets erg technisch wordt kan je een deel naar de Appendix verplaatsen.

Probeer er een lopend verhaal van te maken.

Het is heel handig dit ook weer op te delen nav je deelvragen:

3.3.1 RQ1

3.3.2 RQ2

4 Evaluation

Met een subsectie voor elke deelvraag.

In hoeverre is je vraag beantwoord?

Een mooie graphic/visualisatie is hier heel gewenst.

Hou het kort maar krachtig.

5 Conclusions

Hierin beantwoord je jouw hoofdvraag op basis van het eerder vergaarde bewijs.

5.1 Acknowledgements

Hier kan je bedanken wie je maar wilt.

References

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In *Mining text data* (pp. 163–222). Springer.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188–1196).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

A Slides