**Literature Summary**

1. Classification algorithms

Vikas K Vijayan et al. 2017. A Comprehensive Study of Text Classification Algorithms. *ICACCI,* 1109-1113.

|  |  |
| --- | --- |
| **Abstract:** | The paper discusses important algorithms used in the field of text classification |
| **Research Question/Topic:** | NONE |
| **Method:** | NONE |
| **Results/Evaluation:** | NONE |
| **Limitations:** | NONE |
| **Summary:** | * “Text classification is process of classifying text documents into fixed number of predefined classes” * is used for different tasks like spam filtering or language identification * process: During the text classification process a set of documents {d\_1, d\_2,…..d} are categorized into a set of classes {c\_1,…c\_n}. Thus the task falls into supervised machine learning. * Classification tasks can only include one class, namely single labelled classification or can include multiple classes. Depending on whether it is multi-labeled or single classification different challenges occur during the classification problems. * Different stages in classification: pre-processing, text representation and dimensionality reduction, classification itself and evaluation of the performance * Pre-processing: stop word removal, stemming * Text representation: bag of words vs. string representation * Dimensionality reduction like document frequency, Chi-Square or Term Clustering * Classifier: Rocchio Classifcation, KNN Classifcation, Naïve Bayes classification, SVM Classifcation, Regression Based Classifaction, Neural Network Classification, Rule Based Classifcation, Decision Tree Classifcation |

Colas. F., Brazdil P.2006. Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks.in IFIP International Federation for Informatin Processing, 169-178.

|  |  |
| --- | --- |
| **Abstract:** |  |
| **Research Question/Topic:** | Does SVM outperforms KNN and native Bayes? |
| **Method:** | They used binary classification problem to compare naïve Bayes and KNN and SVM. As evaluation they used 10-fold cross validation. As measurement they choose F\_1 measure. |
| **Results/Evaluation:** | Results indicate that Bayes and knn show overall good performance and are faster than SVM.  For large training sets SVM is not optimal, also it is more complicated than the other approaches. But on the other side good for non-linear problems. |
| **Limitations:** | Only on binary classification tasks. The authors admit that classification problems depend highly on the adopted methodology or on task. |
| **Summary:** | * Naïve Bayes and KNN algorithm are simple and have generally a good performance. * SVM: is a binary classifier if more classes we need a SVM classifier   + Classifies by finding a decision boundary, which maximizes the margin between the data points of both classes * K-nearest neighbor: using a test data point the distance between the test point and all other points is calculated. Then, for each class the distances are summed up. The data point is then assigned to the class, to which is has a closer distance. The distance metric has to be defined beforehand. * Naïve Bayes: Using prior probability of a class, as well as the conditional probability the bayes classifier assign the data point to the class which maximizes the probability (nochmal wo anderes nachschlagen) |

Kowarsi, K et al.2019. Text Classifcation Algorithms: A Survey. *Information,*

|  |  |
| --- | --- |
| **Abstract:** |  |
| **Research Question/Topic:** | Overview of different algorithms, evaluation techniques |
| **Method:** | NONE |
| **Results/Evaluation:** | NONE |
| **Limitations:** | NONE |
| **Summary:** | * Parts of Text Classfication: Feature extraction, dimension reductions (optional), classifier selections and evaluations * Set of documents with X\_1, X\_2 (data points), each of them have a class label, evaluation * Different level 🡪 document level, paragraph level, sentence level, sub-sentence level 🡪 goes in more detailed about how documents can be different than Vikas K Vijayan et al. 2017 * Feature Extraction: TF-ID, TD, Word2Vec, GloVe * Dimensionality Reduction: PCA, LDA, NMF, t-SNE * Classifiers:   + traditional methods like Rocchio   + ensemble-based learning techniques like boosting and bagging (“mainly for text analysis, query learning strategies”)   + simplest classification algorithm: LR, Naïve Bayes   + non-parametric techniques like k-nearest   + SVM   + Decision Tree   + Conditional Random Field   + Random Forest   + Deep Learning approaches * Evaluation:   + Simplest technique: accuracy, but does not work for unbalanced data sets   + Sensitivity   + Specificity   + Precision |

Sebastiani F. 2001. Machine Learning in Automated Text Categorization

|  |  |
| --- | --- |
| **Abstract:** |  |
| **Research Question/Topic:** | Overview + Definition |
| **Method:** | NONE |
| **Results/Evaluation:** | NONE |
| **Limitations:** | NONE |
| **Summary:** | * Single vs. multilabel classification   + Usually, we can just transfer the problem of single-label classification to multi-label classification, by dividing the problem into n binary classification problems   + But this requires that the categories are stochastically independent of each other * Also again a lot of information (same to the other overviews about different methods of classification) |

1. Short classification algorithms

|  |  |
| --- | --- |
| **Abstract:** |  |
| **Research Question/Topic:** | This paper gives an overview about text classification for short text classification |
| **Method:** | NONE |
| **Results/Evaluation:** | NONE |
| **Limitations:** | NONE |
| **Summary:** | * main problem with short text is that the text segments have not enough word-occurrence and no shared context, which is necessary for a qualitative similarity measurement. * Most of the traditional methods are based on word frequency, high word co-occurrence and shared context. Thus those methods are appropriate for classification tasks including short text documents. * Features of short text classification:   + Sparseness: few words, no word co-occurrence or shared context   + Immediacy and Non-standardability: text sent immediately and often nonstandard terms, noise and misspellings. This might not a big problem for the task in this case here, since we have no “classic” short text in the sense of messages or online posts. Also the author mention that short text data often faces imbalanced distribution and problems in labeling of large scale data. The advantage of the training data set used in this paper, is that it comes already labelled. Also, since to the huge amount of data, it is possible to re-scrape the data for a specific class. |
|  | * Since to these challenges of short texts traditional methods like SVM, Bayes or KNN, which are “based on the similarity of term frequency” , often fail to accurately classify. They fail if we do not provide enough labeled information. * “Short text has weaker capacity of semantic expression, which is needed this correlationship. While traditional classification cannot distinguish language fuzziness of natural language, cognates and synonyms all which are abundant in short text” 🡪 fail to achieve good accuracy with for short texts. * Due to the problems short text classification comes with, the author present three methods which are useful for short text classification. First is to reduce the feature dimension by using semantic relationship. He presents the models LSA pLSA, and LDA. Second semi-supervised learning. This does make sense, when we have a lot of unlabelled data. But in my case, this does not matter since we have all labelled data. Last he suggest to use ensemble short text classification. In contrast to single classifiers, ensemble learning method weight weak classifier appropriately, which allows then for getting the weight of each feature. He proposes the following based on other authors   + Dynamic ensemble (for more information: Yan Rui et al. Dynamic Assembly Classification Algorithm for Short Text)   + IR, Voting (Short text classification using few words)   + Domain Knowledge (Feng, Xioa et al Chinese Short Text Classification Based on Domain Knowledge)   + Domain specific features (Bharath Sriram. Short Text Classification in Twitter to Improve Information Filtering) * Evaluation Methods (just explanation, no discussion)   + Accuracy   + Precision and recall   + F-measure   + Macro average and micro average |

Wang Y, Zhou Z. Jin S. et al. 2017

|  |  |
| --- | --- |
| **Abstract:** |  |
| **Research Question/Topic:** | This paper looks at the performance of short text classification, focusing especially on feature selection methods and by using Chinese short text. |
| **Method:** | As preprocessing they used segmentation (but for English not so important) and feature selection. They differentiate between two representations, either as sparse vectors or as dense vectors. As classifiers they choose Naïve Bayes, decision Tree, k-nearest, logistic regression and svm. They applied the big and the small classes of labels (p.4) |
| **Results/Evaluation:** | Results indicate that logistic regression and svm are better than multinomial naïve Bayes classifier using tf-idf/counters features. Small class works better than class, doc2vec does not produce good results at all. Word2vec does not show good result for big class, for small classes seems to work good with LR especially. Again multinomial Bayes shows not good results. |
| **Limitations:** | NONE |
| **Summary:** | * Problems of short text classification   + Lack of information   + Sparsity could be a problem? * Feature selection   + In Feature selection we need to convert the words into vector matrix 🡪 word embedding or distributional model   + Either spare vectors or dense vectors     - Sparse vectors: most elements equal to zero + high length 🡪 computationally expensive     - Dense vectors are much faster for classification     - Spare vector: counter vectorizer or tf-idf     - Dense vector: shorter,       * Single value decomposition       * **Neural language model (state-of the art)**       * Brown clustering     - Neural language model:       * Skip-gram and continuous bag of words (CBOW) (= word2vec) |
|  |  |

Kushbu Khamar 2013

|  |  |
| --- | --- |
| **Abstract:** | Using headlines from Reuters news articles, the author compares SVM, Naïve Bayes and k-nearest neighbor |
| **Research Question/Topic:** | Comparing of algorithms for short texts classification |
| **Method:** | Applied SVM, Naïve Bayes and k-nearest neighbor |
| **Results/Evaluation:** | Results indicate that k-NN shows best accuracy |
| **Limitations:** | not transparent what steps of preprocessing were conducted |
| **Summary:** |  |

Wang et al. 2014, Concept-based short classification

|  |  |
| --- | --- |
| **Abstract:** |  |
| **Research Question/Topic:** | Short text classification using bag-of concepts |
| **Method:** | The authors criticize that BOW has limitations in short texts since they are more spars, noisy and ambiguous than long documents. Instead, the authors propose a bag of concept method where words with similar concepts should be represented as similarly. E.g., “Jeep” and “Honda” by the concept car. They propose a new method: BocSTC. To do they used knowledge bases to learn the concepts. |
| **Results/Evaluation:** |  |
| **Limitations:** | Problem: We need a knowledge database. As the authors say it is more for lightweight applications. |
| **Summary:** | * BOW ignores the semantic and conceptual information * Short texts are noisy, sparse and ambiguous |

Chen, Hu et al. 2019. Deep Short Text Classifcation

|  |  |
| --- | --- |
| **Abstract:** | The authors classify short texts by using attention mechanism and deep learning models with Knowledge powered Attention. Also they use conceptual information. |
| **Research Question/Topic:** | short text classification |
| **Method:** | The authors combine explicit and implicit representation of short texts into a unified deep neural network model and addition they enrich with concepts (same as Wang et. al 2014). Since of the ambiguity of short texts the authors propose Short text classification with Knowledge Powered Attention. |
| **Results/Evaluation:** | There results outperform current state-of the art methods |
| **Limitations:** |  |
| **Summary:** | * short texts are more ambiguous since there is no context information * short texts can be represented as dense vectors or sparse vectors * KPCNN (Wang et al. 2017) * CNN is a classic baseline for text classification (Kim 2014) |

Bouaziz, A et al. 2014 Short Text Classification Using Semantic Random Forest

|  |  |
| --- | --- |
| **Abstract:** |  |
| **Research Question/Topic:** | The authors apply a semantic random forest method to short text for classification |
| **Method:** | In contrast to previous authors they enrich the data not by just short texts, but by the semantic behind each word of the text and the whole text itself. In order to apply the classification, they use Random Forest. For pre-processing they use Latent Dirichlet Allocation (LDA) to derive topics. In a second step they enriched first the short texts with topics which are like each word in the text. Further they enrich the texts with words from the topics (most similar four topics to the whole text). Then they construct with random feature selection the trees. |
| **Results/Evaluation:** | In contrast to classical random forest the authors improved the classification by semantic enrichment |
| **Limitations:** |  |
| **Summary:** | * Short Text: word sparseness, lack of contextual information and often informal sentence expressiveness. * Often used approach enrich data, can produce noise. We have to think about the real improvement before just enriching the data |

Wang F. et al 2014: Concept-based short text classification and ranking

|  |  |
| --- | --- |
| **Abstract:** |  |
| **Research Question/Topic:** | The authors present a different representing format for short text for classification. |
| **Method:** | Instead of using the traditional “bag or words” concept, which is based on the direct appearance of the words in the text, the authors propose a “bag of concepts” approach. In this approach a concept is some sort of domain, whereby words with similar domain become the same representation. For example, Jeep and Honda are represented by car. Advantage: Handle in some extend surface mismatch since it matches at concept level and not at the surface level.  Approach as follows:   * + Bag of Concept for each class (taxonomy knowledge base for converting terms to concepts is needed)   + Then each short text is to be conceptualized with the concepts (“associate with the concepts”) |
| **Results/Evaluation:** |  |
| **Limitations:** |  |
| **Summary:** | * Bag of words results in high dimensionality, problem of surface mismatching * In short text this is even more serious due to the sparsity and shortness of short text * Short texts are noisy, sparse and ambiguous |

1. Works about multi-class classification

|  |  |
| --- | --- |
| **Abstract:** |  |
| **Research Question/Topic:** | introduce a technique called boosting |
| **Method:** | NONE |
| **Results/Evaluation:** | NONE |
| **Limitations:** | NONE |
| **Summary:** | * Basic idea: combine simple categorizations to one highly accurate classification * New and improved family of boosting, based on AdaBoost algorithm * This here focuses more on multi-class classification * Extension of AdaBoost designed especially for multi-class classification 🡪 BoosTexter * In single label classification a set of weights is maintained over the training examples * Actually the algorithm forces the weak learners to focus on the hardly classifying cases * In multiclass problems it is in contrast better to maintain also weights over the labels. By doing so, not only the algorithm forces the hard classifying cases, but also the hard cases labels. |

1. Evaluation
2. TODO

* Challenges for my task:
  + Multi-label classification (especially level 5)
  + Short text classification
* Literatur:
  + Gabrilovich and Markovitch 2007: Computing semantics relatedness using Wikipedia-based explicit semantic analysis
  + Wang et al 2014 Concept based short text classification and ranking
  + Wang and Wang 2016: Combining knowledge with deep convolutional neural networks for short text classification
  + Wang et al. 2017. Combining knowledge with deep convolutional neural networks for short text classification and ranking
  + Kim 2014 Convolutional neural networks for sentence classification
  + Lee J. Y. et al. 2016 Sequential short-text classification with recurrent and convolutional neural networks

|  |  |
| --- | --- |
| **Abstract:** |  |
| **Research Question/Topic:** |  |
| **Method:** |  |
| **Results/Evaluation:** |  |
| **Limitations:** |  |
| **Summary:** |  |