St. Pölten University of Applied Sciences

/medien & digitale technologien

/fh/// st.pölten

# Automated Identification of Information Disorder in Social Media from Multimodal Data

Master Thesis

For attainment of the academic degree of

**Dipl.-Ing.für technisch-wissenschaftliche Berufe**

in the Masters Course Interactive Technologies at St. Pölten University of Applied Sciences, Masterclass Mobile

Submitted by:

**Armin Kirchknopf, BA MA BSc**

it181501

Advisor: FH-Prof. Dipl.-Ing. Mag. Dr. Matthias Zeppelzauer,

Second Advisor: Dipl.-Ing. Djordje Slijepčević, BSc

Sankt Pölten, 30.08.2020

# Declaration

- The attached research paper is my own, original work undertaken in partial fulfillment of my degree.

- I have made no use of sources, materials or assistance other than those which have been openly and fully acknowledged in the text. If any part of another person's work has been quoted, this either appears in inverted commas or (if beyond a few lines) is indented.

- Any direct quotation or source of ideas has been identified in the text by author, date, and page number(s) immediately after such an item, and full details are provided in a reference list at the end of the text.

- I understand that any breach of the fair practice regulations may result in a mark of zero for this research paper and that it could also involve other repercussions.

Date: _____ Signature: _____

# Acknowledgement

At first, I would like to thank my advisor FH-Prof. Dipl.-Ing. Mag. Dr. Matthias Zeppelzauer of the Institute of Creative\Media/Technologies at the University of Applied Sciences St. Pölten. Whenever I needed help he was there. Whether I needed thematic, professional or personal advice. Also his office door (or in Corona times online status) was always open for me.

Secondly, I would like to thank my colleagues, especially Florian Taurer and Djordje Slijepčević, for their continued support and motivation.

Further, I would like to thank Netidee for supporting me with a scholarship, which helped me a lot to finance hardware and offered a lot of new connections into the field of AI research.

And last but not least, I would like to thank my family, but especially my wife Tina Hobel. The last ten years, which we have had the privilege of spending together, have given me a great deal of strength and support. This work would not have been possible without your active support. I love you!

Thank you very much!

# Abstract

Since the Brexit negotiations and the US Election from 2016 the impact of Fake News on public opinion, transported through Social Media, such as Twitter, Facebook, and News are not longer negatable. This thesis tries to develop a method for Information Disorder Detection for the social media platform Reddit based on a publicly available dataset called Fakeddit. For this purpose, a multimodal neural network is designed for tackling this problem. Many other state-of-the-art methods are only using single or dual-modality approaches e.g. text and images. This thesis incorporates up to four different modalities to prove the advantages of a multimodal over a mono-modal approach. The results confirm the superiority of a multimodal approach and improves the detection accuracy by a remarkable amount.

# Kurzfassung

Spätestens seit den Brexit Verhandlungen und den US-Wahlen 2016 sind die Auswirkungen von Fake News auf die öffentliche Meinung nicht mehr verleugbar. Diese Falschinformationen werden großteils über soziale Netzwerke wie Twitter und Facebook verteilt. In dieser Arbeit wird versucht werden, eine Methode zur Erkennung von Fake-Informationen für die Social-Media-Plattform Reddit zu entwickeln, die auf einem öffentlich zugänglichen Datensatz namens Fakeddit basiert. Zu diesem Zweck soll ein multimodales neuronales Netz entworfen werden, um dieses Problem zu lösen. Viele andere State-of-the-Art Methoden beschränken sich auf nur eine oder zwei Modalitäten, z.B. Text und Bilder. In dieser Arbeit werden bis zu vier verschiedene Modalitäten miteinbezogen, um die Vorteile eines multimodalen gegenüber einem monomodalen Ansatz nachzuweisen. Die Ergebnisse bestätigen diese Annahme und zeigen, dass die Kombination verschiedener Modalitäten die Erkennungsgenauigkeit erheblich erhöht.

# Contents

# 1 Introduction

## 1.1 Motivation

At the latest since the US presidential election in 2016, the public has become painfully aware of the influence of "fake news" or in general "information disorder" on public opinion. The topic is highly complex and the evaluation of fake news is a semantically very demanding task. The manual identification of fake news or rumors in a broader sense is a difficult task even for experts. Because of the constantly growing amount of data, the question arises whether information disorder can also be automatically identified and evaluated by data analysis and machine learning, this will be the main goal of this thesis.

## 1.2 Goals

To answer this question, an information disorder detector based on state-of-the-art neural network models from the field of Artificial Intelligence (AI) and Machine Learning (ML), as well as methods of Natural Language Processing (NLP) will be built and evaluated. The aim is to create a method to enable users of social media channels such as Twitter or Reddit to quickly distinguish between fake information and non-fake information.
Based on four available data modalities, namely two different semantic textual information, visual information, and meta data information a concept of a neural network ensemble will be developed. All four data modalities are merged into one large multimodal network which is then used to automatically distinguish between fake and non-fake samples.

## 1.3 Research Questions

The main research questions which are going to be answered in this thesis are:

- Which modality is more meaningful for information disorder detection, two different textual modalities, the visual modality or the meta-data modality?
- To what extent can combined multimodal analysis, as opposed to mono-modal, improve the detection of information disorder in social media data?
- Which network architectures from research are best suited for the multimodal analysis of information disorder?

## 1.4 Research Approach

Based on a Google Scholar search using common terms like "fake news detection", "misinformation detection", "information disorder detection", "rumor detection" and "rumor verification", current papers and surveys are searched for. Besides Google Scholar other platforms like Springer or Elsevier or IEEE are used. As soon as a solid base of representative literature has been established, further literature (e.g. from the Related Work sections) is examined. Based on the current publications of the last five years, a search for publications that have cited them is also conducted.

The second step consists of a more in-depth analysis of selected papers with regard to the methods and data sets used. The focus here is on papers that use multimodal data. At this stage, first indicators can already be collected to answer the third research question, since the evaluation of the literature search is meaningful in this respect.

The third step is to develop an automatic method for meaningful information disorder detection by using state-of-the-art machine learning methods.

## 1.5 Applications

Due to the increasing amount of internet users also the users of online social media websites such as Facebook, Twitter, Instagram or Reddit are steadily increasing. The speed of spreading News (and Fake News) is also becoming faster. For this reason, methods have to be developed to verify these news for credibility. Some of the already online web services are:[1]

| **Fact-checking Website** | **automatic / manual checking** |
|---|---|
| altnews.in | manual |
| Climatefeedback.org | manual |
| factcheck.org | manual |
| factmata.com | automatic (AI) |
| fullfact.org | automatic (AI) |
| hoax-slayer.com | manual |
| hoaxy.iuni.iu.edu | using other online services |
| leadstories.com | manual |
| mediabiasfactcheck.com | manual |
| our.news | automatic (AI) |
| politifact.com | manual |
| snopes.com | manual |
| truthorfiction.com | manual |

*Table 1.1.* Excerpt of available fact-checking websites

As shown in table 1.1 many of the provided websites are performing fact checking manually by doing good journalism. But this is very expensive and time consuming and can be extended by the use of artificial intelligence. This extension is necessary due to the tremendous speed of spreading misinformation. Today, however, only a few platforms employ AI for information verification.

[1] All retrieved on 11.07.2020.

## 1.6 Structure of the thesis

The following thesis is structured as follows. Chapter 2 gives an overview about existing technologies in the field of mono and multimodal information disorder analysis. It starts with a definition of terms, what is currently summarized under the terms e.g. information disorder, misinformation, rumor, spam and Fake News. Afterwards a deep dive into the state-of-the-art is made structured by the modality text, visual and meta. Followed by an overview about the currently used datasets in this field of application. The chapter will be closed by a in-depth discussion about the proposed methods, their advantages and disadvantages and open challenges.

Chapter 3 provides an detailed description of which parts the experimental part of this thesis is build on. The chapter starts with an introduction to the four modalities chosen and where they came from. Furthermore also the possible output is discussed. Afterwards several possibilities about pre-processing each of these modalities is discussed. Section 3.3 an analysis about the used components is provided, followed by a detailed description how every component is implemented and which special components were chosen to handle the challenge of detecting information disorder.

Chapter 4 presents the chosen dataset, all experimental setups and all the results of the experiments carried out and discusses them modality for modality. Finally chapter 5 concludes the whole thesis, presents open questions and future work.

# 2 Background & Related Work

## 2.1 Terminology

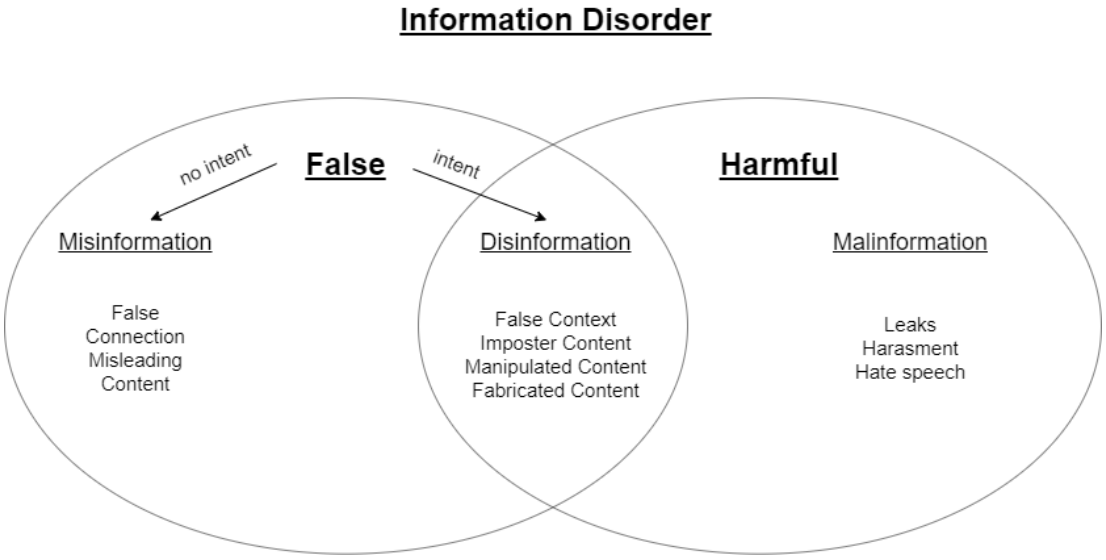


*Figure 2.1.* Conceptual Framework of information disorder, combined image of (Wardle and Derakhshan 2017) and (Kumar and Shah 2018)

The term Fake News is nowadays used in many different cases. It does not describe all facets of the whole complex topic of information disorder. A more general term that can be taken is "information disorder" which can be applied to all historical impacts of mis-, dis-, and malinformation, that we currently know and classify contemporary mis- and disinformation as Wardle and Derakhshan (2017) as "information pollution" at a global scale. An overview and to visualize the whole concept about all three types see furthermore figure 2.1, that will be described in the rest of this section. In general these types can be distinguished either by intent, mis- and disinformation, or if the content is even harmful (malinformation).

### 2.1.1 Misinformation

Misinformation is information which are created without the intent of misleading users otherwise this information is called disinformation if there intent was to mislead (Kumar and Shah 2018). Moreover, misinformation can occur either inadvertently, for example through ignorance, or through misunderstanding of facts. Often misinformation is also distributed quickly through ignorance via blogs or other services such as Facebook or Twitter. But even very trivial facts can be unintentionally distorted by different interpretations or knowledge levels of the people involved.

### 2.1.2 Disinformation

Disinformation in general, are information with an intent to change the public opinion (Fallis 2014; Hernon 1995; Kumar and Shah 2018). A very good example for this kind of information are the Brexit Votum and the US election 2016 (J. Kim et al. 2018), where the public opionion was lead to a certain decision which had a great impact on the political landscape of the world. The problem behind is, as stated in (Kumar and Shah 2018), to understand the motives for the disinformation and the motive for the generation for the deception. Both questions do not necessarily have to be answered with the same statement or intent. Rumours, for example, also fall into this category.

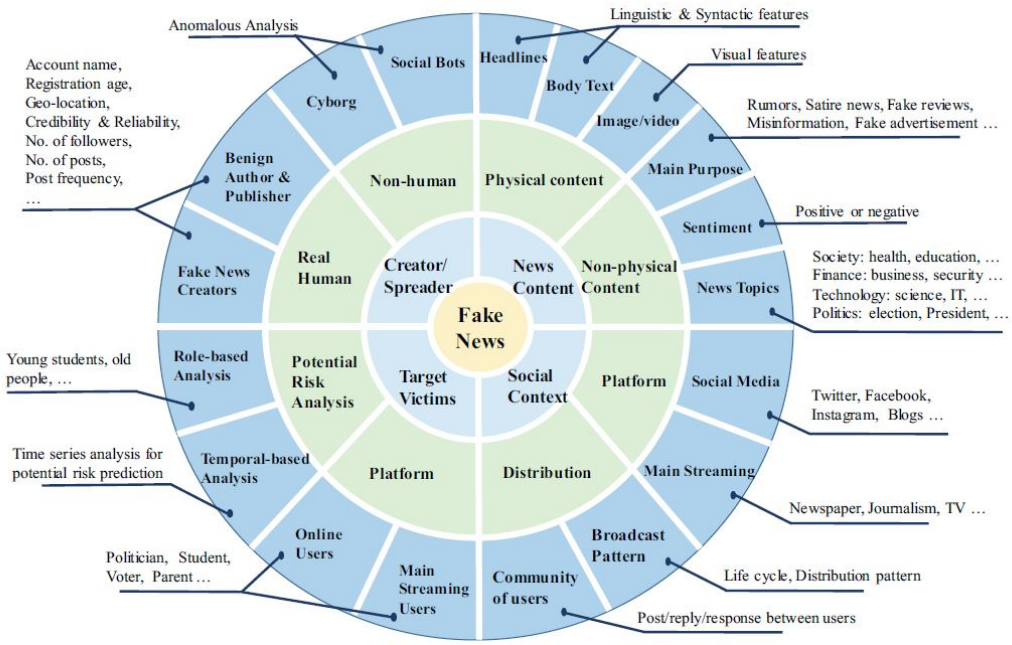


*Figure 2.2.* Fake News and related components, originally by (Zhang and Ghorbani 2020)

One type of Disinformation is Fake News. Fake News is news which intent is to mislead the public opinion so they fall into this category of information. As seen in figure 2.2 the whole problem can not being described only by one term. Fake News, like all news, have to be written or generated. This can be done by humans or computers. The next component is the target of Fake News. These can be a whole platform such as Facebook or Twitter for example. The fake needs to be written for one ore more specific target groups, this is summarized under the term "Potential Risk Analysis". The generated content can be divided in real content such as text or images, but also into non physical content such as purpose or sentiment. The social context part describes how the fakes are distribute, for example via a platform, this includes again an analysis of the target group. The last (blue) components describes the components before but in a more fine grained way to specify each component in more detail. To summon up, Fake News is just a term which describes a high-level concept for distributing manipulated content to one or more target groups on specific platforms with specific sentiment and language (Zhang and Ghorbani 2020).

In the worst-case Fake News can lead to a gunshot[1]. A great impact of spreading Fake News are so-called social bots, which are automated programs on social media services, such as Facebook and Twitter which are spreading misleading content (E. Ferrara et al. 2016; Forelle et al. 2015; P. N. Howard and Kollanyi 2016; Shao, Ciampaglia, Varol, Flammini, et al. 2017; Shao, Ciampaglia, Varol, Yang, et al. 2018). Their intent is to lead a discussion into a certain direction, all controlled by one person, a so-called lone-wolve (Kumar and Shah 2018; Vosoughi et al. 2018). So just one person can be responsible for spreading information disorder with the intent to mislead the public opinion.

To summon up, Fake News is news which is intentionally and verifiable false and may mislead readers (Allcott and Gentzkow 2017).

### 2.1.3 Malinformation

Malinformation is a kind of information, created with the intent to harm people or organisations. The interesting part is that the information used, is true information and not generated information. So truthful information, for example court records or classified material is published. One example is the leak of emails of Macron just before the run-off vote in 2017.[2] The information within the emails was real, but the handling of the case included false information and harmed Macron's campaign.

[1] As happened in 2016: https://wapo.st/2FVhoEZ, retrieved on 12.07.2020.

[2] For more information about the email leak see: https://www.nytimes.com/2017/05/08/world/europe/macron-hacking-attack-france.html, retrieved on 24.08.2020.

### 2.1.4 Fine-grained categorization

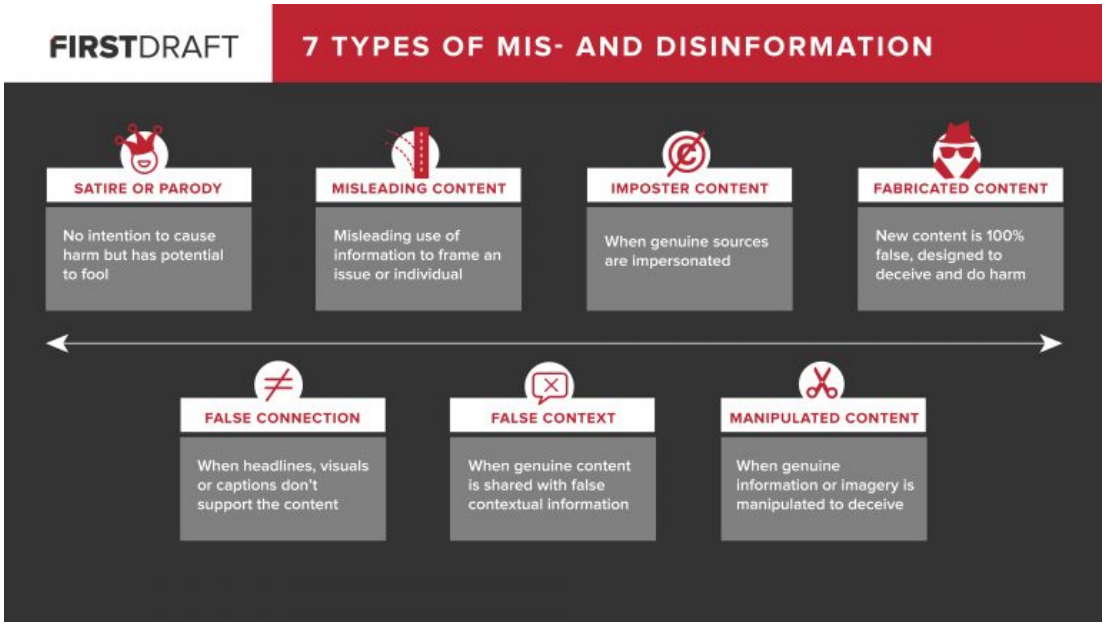


*Figure 2.3.* Types of information disorder, as shown in (Wardle 2017; Wardle and Derakhshan 2017).

A more fine-grained categorization of mis-, and disinformation is proposed by Wardle and Derakhshan (2017). It is often not possible to draw a hard line between mis-, and disinformation. One aspect of this kind of information is the intent. As seen in figure 2.3 on the one hand, "Satire or Parody" on the left side has no intention to cause harm but can fool people, but one the other hand "Fabricated Content" is designed and has the intent to be designed and mislead people. So on the left hand side of the figure several types of misinformation and on the right hand side several types of disinformation are shown. In between there are several gradations of the respective extremes. It is important to note that it is a continuum, i.e. there is no sharp or clear border between the individual types of mis-, and disinformation. The 6-way label of the chosen dataset, which is introduced in section 4.1, has the possibility to train a method for distinguish between most of these types.

### 2.1.5 Focus of this thesis

This thesis will mainly focus on information disorder as introduced by Wardle and Derakhshan (2017), which is defined as the combination of the three types of information introduced before: mis-, dis-, and malinformation as seen in figure 2.1.

Furthermore information disorder has three different phases, or life cycle:

- Creation - When the message has been created
- Production - When the message has been converted into a media product
- Distribution - The message is made public

Similar to the lifecycle following three elements of information disorder can be defined:

- Agent - The person/program/bot that created, produces and distributed the message
- Message - Important to find out is: The type of message, format of the message and occurring characteristics
- Interpreter - When the message was read by the target group, how is it interpreted, which actions are performed afterwards.

To summon up, information disorder is an interdisciplinary concept that tries to describe the universe of fake information generation, distribution and social as well as historical impact on our current society (Wardle and Derakhshan 2017). This thesis will focus on the detection of fake and non-fake information, summarized under the term information disorder.

## 2.2 State-of-the-Art

This section surveys current state-of-the-art methods for information disorder detection for different individual data modalities and multiple modalities. Table 2.1 should provide an overview about state-of-the-art papers and their used dataset modalities. This table by no means claims to be complete, but offers a representative extract of the current research situation. Within the last five years more and more researchers dealt with the topic of NLP and developed a new set of methods for text and language processing. Most of the research was done by using the textual modality. Fusing different information sources such as, for example, text and visual information is a relatively new topic. The literature research focused on a selection of high-quality paper from the last few years. All papers are by renowned and much-cited authors and this state-of-the-art analysis should provide a good overview about the topic of information disorder detection.

| Author | text | visual | meta |
|---|---|---|---|
| Ma, Gao, P. Mitra, et al. (2016) | ✓ | | |
| Ma, Gao, and Wong (2017) | ✓ | | |
| Ma, Gao, and Wong (2018) | ✓ | | |
| Mohtarami et al. (2018) | ✓ | | |
| Lago et al. (2019) | | ✓ | |
| Singhal et al. (2019) | ✓ | ✓ | |
| Y. Wang et al. (2018) | ✓ | ✓ | |
| Nakamura et al. (2020) | ✓ | ✓ | |
| Dong et al. (2018) | ✓ | | ✓ |
| Zubiaga, Liakata, and Procter (2017) | ✓ | | ✓ |
| Ruchansky et al. (2017) | ✓ | | ✓ |
| Cui et al. (2019) | ✓ | ✓ | ✓ |
| Jin et al. (2017) | ✓ | ✓ | ✓ |

*Table 2.1.* Overview about authors and used modalities in the proposed methods.

### 2.2.1 Mono-modal approaches incorporating textual information

Ma, Gao, and Wong (2018) proposed a rumor detection system with the ability to detect rumors at a very early stage by using tree-structured recursive bottom-up (BU-RvNN) and top-down (TD-RvNN) neural networks. Each word of each claim is modelled as a tree architecture and processed by an LSTM and gate recurrent units (GRU) (Cho et al. 2014). The GRU should model the interaction between each tree node maintaining the structural integrity of a sentence or a claim in general. So the advantage of the proposed method lies in learning more meaningful representations of sentences and does not need to perform expensive pairwise comparisons of words. The used datasets Twitter15 and Twitter16 are smaller datasets that consists out of around 1.2k samples each and were introduced by Ma, Gao, and Wong (2017). The proposed method BU-RvNN reached an accuracy of 70.8% and the TD-RvNN reached 72.3% on the Twitter15 dataset. Furthermore the BU-RvNN reached an accuracy of 71.8% and the TD-RvNN reached 73.7% on the Twitter16 dataset for the task of rumor detection.

Mohtarami et al. (2018) proposed a stance verification system that uses a CNN in combination with LSTMs. The aim was to verify if a given document agrees, disagrees, discusses, or is unrelated to a certain claim. The used dataset "Fake News Challenge" will be introduced in section 2.3. The word embeddings were generated by using GloVe (Pennington et al. 2014). The CNN connects neighbouring word units to improve the very promising results. Due to the unbalanced dataset randomly selected samples of all classes allows training on the same number of samples for each class. The method is also capable to extract meaningful snippets which can help experts for further investigation and explainability of the method. The best of the three proposed methods (sMemNN with TF) performed with an accuracy of 88.57% on the FakeNewsChallenge dataset.

Kochkina et al. (2017) proposed a rumour veracity and rumour stance classification system which uses a Branch-LSTM Model (Hochreiter and Schmidhuber 1997). The used dataset will be described in section 2.3. The method uses only textual information and information calculated out of the tweets for example the count of negation words or swear words. The text is pre-processed with standard NLP methods such as removing stop words, lower case, and tokenization. The proposed method is capable to learn features from the tweet representation together with the other features at each time step. The proposed method performed with an accuracy of 78.4% on the test set of PHEME dataset for the task of rumor veracity and stance classification.

### 2.2.2 Mono-modal approaches incorporating visual information

Lago et al. (2019) proposed an image forensic technique-based approach. Textual analysis, as stated, has only a complementary effect on the classification. By applying image forensic methods such as color filter array (P. Ferrara et al. 2012) or jpeg ghosts analysis (Farid 2009; Zach et al. 2012). A similar tool is called Splicebuster (Cozzolino et al. 2015), it is possible to find edited or manipulated images. The used datasets "image verification corpus" and "BuzzFeedNews" will be described in section 2.3. A third dataset was created by crawling news websites by the authors themselves. It could be shown that classical image verification methods, as described above, work on almost the same level as Splicebuster and machine learning methods like Random Forest Classifier (RFC) or Logistic Regression Classifier (LRC). The described results refer only to the image modality and are not representing combinations of image and text.

The proposed methods RFC reached an accuracy of 74% on the "image-verification-corpus" dataset by using only image forensic technologies as stated in the paper. The same accuracy was reached by the combination of RFC and the Splicebuster method. Only the RFC with CNN method performed around 73%. A similar picture is provided if the LRC is taken into account. The accuracy of LRC in combination with the image forensic method reached around 75% accuracy on the dataset. The same accuracy was achieved by combining LRC with Splicebuster. Again only the LRC + CNN method was performing around 74%.

The RFC in combination with the image forensic methods achieved around 75% accuracy on the BuzzFeedNews dataset. The RFC-Splicebuster method performed with 70% less well on this dataset. Here the RFC-CNN method performed with 73% much better. The LRC in combination with the image forensic methods performed around 77%. The LRC-CNN method with around 78% on the BuzzFeedNews dataset was performing a little bit better. Only the LRC-Splicebuster method performed with 76% slightly worse than the other methods on this dataset.

The RFC in combination with the image forensic methods achieved, equally to the CNN method, around 92% on the self-crawled CrawlerNews dataset. Only the Splicebuster method performed less well with around 88%. Utilizing the LRC all three methods performed equally with around 93% on the CrawlerNews dataset.

### 2.2.3 Multimodal approaches incorporating textual and visual information

Nakamura et al. (2019) and Nakamura et al. (2020) proposed a multimodal network for Fake News Detection by using the textual and visual modality. The used dataset Fakkedit will be the evaluation dataset for this thesis and an in-depth analysis can be bound in section 4.1. For the textual modality two different methods namely BERT[3] and InferSent (Conneau et al. 2017) were chosen for the evaluation. For the visual modality three different commonly used image classification networks ResNet50 (He et al. 2016a), VGG16 (Simonyan and Zisserman 2014) and EfficientNet (Tan and Q. V. Le 2019) were chosen for the feature extraction. These networks were used to extract features from the dataset which are used as input for the classification head. This consists out of a fusion layer which fuses the features by following strategies: Add, Average, Concatenate and Maximum.[4]

The proposed method, combining textual and visual information performed with around 90% validation and test accuracy very well on the Fakeddit dataset regarding the 2-way problem. The same accuracy was achieved on the 3-way problem. Only the accuracy on the most complex 6-way problem dropped to around 86% on both sets.

EANN or event adversarial neural network proposed by Y. Wang et al. (2018) is a multimodal network ensemble that utilizes textual and visual information. The advantage above other approaches is that EANN can extract event invariant features. This knowledge can be easily transferred between already known events to unknown events. The textual information is processed by a Text-CNN (Y. Kim 2014) and a VGG-19 for the visual modality. The used dataset "image-verification-corpus" will be introduced in section 2.3. The second Weibo dataset is identical to the dataset used by Jin et al. (2017).

The EANN approach achieved an accuracy of 71.5% on the image-verification-corpus dataset and 82.7% accuracy on the self-collected Weibo dataset. The authors could show that using an event discriminator for fake news detection can improve the accuracy of such systems.

[3] Which will be discussed in section 3.1.1.
[4] https://keras.io/api/layers/merging_layers/, retrieved on 21.07.2020.

Spotfake (Singhal et al. 2019) is a multimodal fake news detection system which utilizes BERT (Devlin et al. 2018), which will be described in more detail in section 3.1.1 for text processing and VGG-19 (Simonyan and Zisserman 2014), pretrained on Imagenet (Deng et al. 2009) for the visual modality. One of the used dataset is the "image-verification corpus" which will be introduced in section 2.3. The second dataset from the Chinese social media website Weibo[5] is also taken into account. The extracted features from BERT and the VGG-19 model are concatenated. The results of the Spotfake system with an accuracy of 77.77% on the "image-verification-corpus" dataset and 89.23% on the Weibo dataset, which were also used by the EANN approach showed that the Spotfake system performs better than the EANN approach(Y. Wang et al. 2018).

[5] weibo.com, retrieved on 13.07.2020.

### 2.2.4 Multimodal approaches incorporating textual and meta information

Dong et al. (2018) proposed a unified attention model with latent representations by combining textual information and users meta data. The used data comes from the "LIAR" dataset which will be introduced in section 2.3 and from the BuzzFeed dataset which will be also described in section 2.3. The strength of the proposed hybrid models lies in the ability to fuse information consisting out of different kinds of information, such as textual and meta data. The RNN component can capture contextual information at a high level but cannot hold information about the prominent parts of the textual modality. To overcome this disadvantage the method utilizes bidirectional gated recurrent units (GRUs) (Cho et al. 2014) to extract features from the data. All the extracted information are fused in an attention matrix which fuses the features from both modalities. The results showed that a multimodal approach succeeds above monomodal settings by achieving an accuracy of 82.83% on the LIAR dataset and 83.84% on the BuzzFeedNews dataset.

Zubiaga, Liakata, and Procter (2017) proposes a rumor detection system by utilizing Conditional Random Fields. The dataset (PHEME) will be discussed later in section 2.3. Each tweet and response on it has been modelled as a chain to take advantage of the CRF method which can also consider neighbours to its decision. The data chain consists out of the text of the tweet together with relevant meta data (social context) of the user which posted it. This method proposed, outperforms other classifiers such as SVM, Random Forest, and others on the PHEME dataset. The best classifier, measured by the F1 score is the conditional random field classifier with a F1 score of 0.606 on the task of rumor detection on the PHEME dataset.

Ruchansky et al. (2017) proposed a Fake News Detection System called CSI - Capture Score and Integrate. The used network architectures utilized a LSTM for learning a representation of the text, preprocessed by doc2vec (Q. Le and Mikolov 2014), an extension of word2vec (Mikolov et al. 2013) and user activity information. The second module "Score" is a neural network fed by a user graph to extract a score for each different user. The "Integrate module" combines both information and classifies both separated learned modality into Fake or no Fake. The used dataset is initially created by Ma, Gao, P. Mitra, et al. (2016) but is currently not available for download. The proposed method achieved 89.2% on the Twitter dataset and 95.3% on the Weibo dataset introduced in Ma, Gao, P. Mitra, et al. (2016).

### 2.2.5 Multimodal approaches incorporating three modalities

SAME or Sentiment aware multi-modal embedding for detecting Fake News (Cui et al. 2019) is a mulitmodal Fake News Detection system which uses GloVe embedded feature vectors (Pennington et al. 2014), VGGNet encoded image features and one-hot encoded meta data such as sources, keywords, and other related information. All the collected information is connected to an adversarial network. The used dataset "FakeNewsNet" will be introduced in section 2.3. The method achieved very promising results on t the Politifact dataset with and Macro F1 score of about 77.24 and 80.42 on the GossipCop dataset and showed clearly the benefits of a Deep Learning based approach over a "classic" machine learning approach such as KNN or SVM.

Jin et al. (2017) proposed a multimodal end-to-end recurrent neural network for rumour detection in microblogs by utilizing LSTMs for textual and meta data information, such as hash-tag, mentions, retweets, and text semantic features, and a VGG-19 network for the visual modality. The authors created its own dataset by crawling the Chinese social media website Weibo between May 2012 to January 2016. The second dataset the "image-verification-corpus" will be introduced in section 2.3. The achieved accuracy around 78.8% on the collected Weibo dataset and 68.2% on the "image-verification-corpus" shows the improvement on the task of rumor detection by fusing information out of different modalities.

## 2.3 Datasets

This section will provide an overview of currently used datasets for information disorder detection. The provided list in table 2.2 is by no means complete and all-encompassing but should give a good entry point into detecting information disorder. The sorting by modality should provide a quick overview about the most commonly used modalities and datasets within the last years.

| Dataset name | Samples | Source | Modality | Url |
|---|---|---|---|---|
| Breaking! | 700 | BS - Detector | text | - |
| COVID-19 Infodemic | 1100 | self-collected | text | https://bit.ly/30apgtN |
| Credbank | 60 mio | crowd-sourced | text | https://bit.ly/3fcNCrc |
| Fake News Challenge | 525k | news websites | text | https://bit.ly/3jUOvZ2 |
| FakeNewsCorpus | 9.4 mio | opensources.co | text | https://bit.ly/3jXFoGG |
| FA-KES | 804 | 15 news websites | text | https://bit.ly/2PbgDcc |
| FEVER | 185k | Wikipedia | text | https://bit.ly/3jTNf8d |
| NELA-GT-2019 | 1.12 mio | 260 news websites | text | https://bit.ly/2DoyU2P |
| PHEME | 15k | Twitter | text | https://bit.ly/312DsEj |
| Twitter15 | 842 | Twitter | text | - |
| Twitter16 | 990 | Twitter | text | - |
| BuzzFace | 2263 / 1.6 mio comments | Facebook | text, image | https://bit.ly/39NzLqc |
| The PS-Battles Dataset | 100k | Reddit | image, meta | https://bit.ly/3fhbYQm |
| FakeNewsNet | 600k | Politifact, GossipCop, Twitter | text, meta | https://bit.ly/3f7vPBC |
| Liar | 13k | Politifact | text, meta | https://bit.ly/3geQJQE |
| NELA2017 | 136k | 92 news websites | text, meta | https://bit.ly/3jVajUg |
| NELA-GT-2018 | 713k | 194 news websites | text, meta | https://bit.ly/3hOjpQO |
| Some Like it Hoax | 15.5k | Facebook | text, meta | https://bit.ly/2P6orfj |
| BuzzFeed 2016 | 2.2k | different news outlets | text, image, meta | https://bit.ly/30aAtKW |
| Fakeddit | 1 mio | Reddit | text, image, meta | https://bit.ly/2Pm9UMJ |
| Image-verification-corpus | 18k | Twitter | text, image, meta | https://bit.ly/30ZNaHv |

*Table 2.2.* Overview about commonly used datasets and where to find them.

**Breaking!** The "Breaking!" dataset consists out of 700 news articles from USA politics from August to November 2016. It incorporates only text samples. It is built out of well known fake articles from the Stanford dataset (Allcott and Gentzkow 2017) in combination with manually labelled fake articles from the famous BS-Detector. [6]. The labelling is done manually by applying two different label categories, primary label - false, partial truth, and opinions/commentary presented as facts and secondary label - fake (from the original Stanford dataset) and questionable (from the original Kaggle dataset). The trained model consists of a bi-directional LSTM in combination with a 1-D CNN model and performed very well on the dataset (Pathak and Srihari 2019).

[6] The original dataset does not exist anymore but has been republished and extended here: https://github.com/thiagovas/bs-detector-dataset The BS-Detector plugin can be found here: https://github.com/selfagency/bs-detector.

**BuzzFace** The "BuzzFace" dataset consists out of 1.6 millions of text samples and related images (if available) from nine news outlets[7] during September 2016, also in the time of the US elections.[8] The first step of quality control was, that only sources were taken that have the "verified" status of Facebook. The whole dataset is then manually labeled by BuzzFeed journalists into four different label categories namely: mostly true, mostly false, a mixture of true and false, and no factual content. No specific baseline method was proposed by the dataset authors (Santia and Williams 2018).

**BuzzFeed** The "Buzzfeed" news dataset[9] is a fact-checking dataset, mainly sourced from Facebook. It consists of around 2.2k samples with links to Facebook posts from Fake News- and non-Fake News websites and also provides several meta data samples. Additional information can be found by accessing the Facebook API.

**COVID-19 Infodemic** The "Covid-19 infodemic" dataset consists out of 1100 news articles from 88 news sources and social network posts from various sites such as Facebook. The provided labels are True or Fake. The balanced dataset was manually labeled.[10] No specific baseline or method was proposed.

**Credbank** The "Credbank" dataset consists out of more than 60 million tweets grouped into 1049 news events. The main source for this dataset is Twitter. The annotation of the dataset was made by utilizing Amazons Mechanical Turk service.[11] For each sample several persons could rate and algorithms were used to sort the ratings into the five possible points on the 5-point Likert scale namely: certainly inaccurate, probably inaccurate, uncertain (doubtful), probably accurate, and certainly accurate (T. Mitra and Gilbert 2015).

**Fake News Challenge** The "Fake News Challenge" dataset consists of around 525k samples collected from different news websites belonging to around 300 topics with 5 - 20 articles. Each sample is labeled into four different categories: agree, discuss, disagree, and unrelated. For now, there is no overview or in-depth analysis of the dataset itself published (Pomerleau and Rao 2017).

[7] See furthermore: https://www.facebook.com/journalismproject/indexing-facebook-news-pages-ad-archive, retrieved on 10.07.2020.

[8] The data was collected by using the Facebook Graph API https://developers.facebook.com/docs/graph-api/, retrieved on 10.07.2020.

[9] https://www.buzzfeednews.com/article/craigsilverman/partisan-fb-pages-analysis, retrieved on 13.07.2020.

[10] See furthermore: https://towardsdatascience.com/explore-covid-19-infodemic-2d1ceaae2306, retrieved on 10.07.2020.

[11] https://www.mturk.com/, retrieved on 10.07.2020.

**Fakeddit** This dataset will be described in detail in section 4.1 because this thesis will be using it for developing a multimodal method for information disorder detection.

**Fake News Corpus** The "Fake News Corpus" dataset consists currently out of 9.4 million samples from 1001 web-domains, mainly from a curated list of websites which can be found here: https://github.com/OpenSourcesGroup/opensources.[12] Each sample of the dataset is labelled to one of eleven types of misinformation, such as fake news, satire, extreme bias, conspiracy theory, state news, junk science, hate news, clickbait, proceed with caution, political, and credible. A baseline implementation can be found on GitHub.[13]

**Fake News Net** The "Fake News Net" dataset consists out of 600k samples from the news websites GossipCop[14] and Politifact, both famous fact-checking websites. From the labeled news articles from the before mentioned websites related Tweets from Twitter were collected, together with user and other meta data. The meta and user data were used for automatic bot detection (Davis et al. 2016). The creators of the dataset also proposed different baseline methods (Shu et al. 2019).

**FA-KES** The "FA-KES" dataset consists out of 804 news articles in the field of the Syrian War. The annotation was supported by crowd-sourcing manually labelled. Additional features were collected and added to the dataset including when did this event happen and/or where did it happen (Salem et al. 2019).

**FEVER** The "FEVER" - Fact Extraction and VERification dataset consists out of 185k samples for the task of claim verification, a subpart of misinformation detection. The source are Wikipedia sentences which have been manually altered and classified into three different label classes namely: supported, refuted, and not enough info. A baseline approach consisting out of a Neural Network utilizing classic text approaches such as TF-IDF for vector similarity across different documents and Natural Language Processing (NLTK) was proposed by (Thorne et al. 2018).

[12] The website http://www.opensources.com/ is offline but still available on the web.archive: https://web.archive.org/web/20190801000000*/http://www.opensources.co/, retrieved on 10.07.2020.
[13] https://github.com/several27/FakeNewsRecognition, retrieved on 10.07.2020.
[14] https://www.gossipcop.com/, retrieved on 11.07.2020.

**Image-Verification-Corpus** The "image-verification-corpus" dataset consists out of 17 different events with around 400 images including 18k tweets of different languages with additional meta data such as user data. For labelling the original news articles were taken into account. If the image was on this reputable news website, the specific image was labeled with True or if not with Fake, the same approach was used for the tweets. A basic approach and baseline were also proposed by the creators of the dataset (Boididou et al. 2018).

**LIAR** The "LIAR" dataset consists out of around 13k manually labeled dataset of short statements of the fact-checking website Politifacts.[15] Their approach uses a combination of word embedding models and meta data which is fed via a ConvNet and a Bi-directional LSTM (W. Y. Wang 2017).

**NELA2017** The "NELA2017" or NEws LAndscape dataset consists out of around 136k samples collected from April 2017 till Oktober 2017 from 92 news sources. The main sources are reliable news websites along with different others such as fake news websites. The content includes articles from the USA political landscape. The dataset consists not only out of the text and body of each news article but also out of a lot of more computed features such as Part-of-Speech (POS) sentiment analysis, Facebook API engagement, readability measure, moral and many more. A full list of features and a detailed analysis of the dataset can be found in (Horne et al. 2018).

**NELA-GT-2018** The "NELA-GT-2018" dataset consists out of around 713k samples collected from February 2018 to November 2018 from around 194 news and media sources. It extends the "NELA2017" dataset by collecting related and similar news articles. The ground truth data is collected from eight different sites below are for example Wikipedia, OpenSources, and Politifact (Nørregaard et al. 2019).

**NELA-GT-2019** The "NELA-GT-2019" dataset consists out of around 1.12 million samples collected from 01.01.2019 till 31.12.2019 from around 260 news sources. It also extends the "NELA-GT-2018" dataset by again collecting around 400k more news articles. The labelling process is similar to the previous dataset (Gruppi et al. 2020).

[15] https://www.politifact.com/, retrieved on 11.07.2020.

**PHEME** The "PHEME" dataset consists out of around 5k tweets grouped into 330 rumor threads. All the collected events are manually labelled by journalists. The topics vary from the Ottawa Shooting over Charlie Hebdo to the Germanwings crash. The labels consist out of rumour threads and non-rumour threads. Within each thread, the conversation tweets are labeled into three types namely: support and response type, certainty, and evidentiality with regarding subtypes to allow a fine-grained analysis of the news events. It consists out of two types of information: Content-based features, which are for example word vectors or part-of-speech tags, but also social features such as tweet count or follow ratio or age of the tweet owner. A detailed analysis can be found in the paper of the creators (Zubiaga, Liakata, Procter, et al. 2016). A baseline method with CRF has been introduced by (Buntain and Golbeck 2017; Zubiaga, Liakata, and Procter 2017; Zubiaga, Liakata, Procter, et al. 2016).

**Some like it Hoax** The "Some like it Hoax" dataset consists out of around 15,5k Facebook posts and around 900k users. It was collected from July to December 2016 through the Facebook Graph API. The sources consist out of 14 conspiracy- and 18 scientific pages with more than 2.3 million likes. The dataset has textual and meta information, such as the likes related to each post. The main research question was to classifiy posts into hoax and non-hoax just by looking at the likes of each post. The dataset creators proposed a baseline method with remarkable results exceeding 99% (Tacchini et al. 2017).

**The PS-Battles Dataset** "The PS-Battles Dataset" consists out of around 100k images grouped into around 11k subsets sourced mainly from Reddit, Imgur[16] and a various number of smaller file sharing hosts. Not only the original, unmanipulated image but also a varying number of manipulated images in different image formats such as jpg and png are contained in the dataset. Next to the visual modality also meta data is collected are available for example the username of the post owner, community score, and many more (Heller et al. 2018).

**Twitter15** The "Twitter15" dataset consists out of around 421 true and 421 rumour stories from the social media platform Twitter, which were collected in March 2015. The dataset contains only out of textual information. The labels were added by verifying the stories manually (X. Liu et al. 2015). At the time of writing this thesis, the dataset was not available anymore.

[16] www.imgur.com, retrieved on 18.07.2020.

**Twitter16** The "Twitter16" dataset consists out of around 498 rumour and 494 non-rumour stories from the social media platform Twitter, which were collected between March and December 2016 and is similar to the "Twitter15" dataset regarding labeling by using the fact-checking website snopes.com (Ma, Gao, P. Mitra, et al. 2016). At the time of writing this thesis, the dataset was not available anymore.

## 2.4 Discussion

As described in the previous section many different approaches are developed for tackling the broad field of information disorder detection, mis- and disinformation detection, rumor verification, and Fake News Detection. Many of the proposed methods are only considering one modality such as suggested by Cho et al. (2014), Kochkina et al. (2017), Lago et al. (2019), Ma, Gao, and Wong (2018), and Mohtarami et al. (2018). The challenges in information disorder detection are multilayered. On the one hand, the detection should be possible at a very early stage (Ma, Gao, and Wong 2018) but this is not always possible due to the lack of appropriate datasets and the multifaceted nature of rumours. On the other hand an early stage detection is often not reliable because a rumor for example can be true till facts are available which prove it either true or wrong. Another problem is, that there are various social media platforms on the internet, such as Facebook, Twitter and Reddit (and many more ...). The mis,- dis,- and malinformation on different platforms has also different natures and facets and current models are not capable to handle information from different sources. Nakamura et al. (2019), Nakamura et al. (2020), Singhal et al. (2019), and Y. Wang et al. (2018) proposed different kinds of multimodal network architectures by fusing textual and visual information. The used components for each modality vary a lot.

On the one hand, Nakamura et al. (2019) and Nakamura et al. (2020) utilize InferSent (Conneau et al. 2017) and BERT (Devlin et al. 2018) for the textual modality and on the other hand (Y. Wang et al. 2018) use a Text-CNN (Y. Kim 2014) and (Cui et al. 2019) utilizes Global Vector (GloVe) (Pennington et al. 2014) or doc2vec (Q. Le and Mikolov 2014) and word2vev (Mikolov et al. 2013), respectively. Most of the methods were using word embeddings such as BERT which is one of the most powerful tools for text processing.

For the visual modality classical approaches such as ResNet, VGG16, VGG19 (Simonyan and Zisserman 2014) or MobileNet (Nakamura et al. 2019; Nakamura et al. 2020) are taken into account. All these networks were pretrained on the ImageNet dataset to benefit of the learned features (Deng et al. 2009). Lago et al. (2019) proposed and evaluated different methods for finding manipulated images within a dataset by using image forensic methods together with Splicebuster (Cozzolino et al. 2015), a tool which learns image manipulation with no prior knowledge.

Meta-data such as related user data or retweets and many more are a useful source of information. Dong et al. (2018) proposed a unified attention model that can learn representations out of text and users meta data. It could be shown that utilizing gated recurrent units (GRUs) (Cho et al. 2014) for extracting features from multi modal data were very useful in detecting information disorder. Another approach by using an LSTM was undertaken by (Ruchansky et al. 2017) which features were extracted by modeling the dataset inside a graph. But also more "traditional" approaches are used within information disorder detection such as Conditional Random Fields (CRF) by (Zubiaga, Liakata, and Procter 2017) which outperform SVMs and Random Forest classifiers.

The concept fusing feature representations from different modalities is often advantageous than relying on only one modality (Cui et al. 2019; Jin et al. 2017; Nakamura et al. 2020). The key for taking advantage out of fusing different modalities is the correct fusing strategy. But this depends on the availability of each of the modalities. Nakamura et al. (2020) for example evaluated different fusing strategies such as concatenation, addition or maximum of each input feature representation. Cui et al. (2019) fused the features into an attention matrix and calculate the output based on this. Jin et al. (2017) also concatenates the output before feeding the new representation into a LSTM network.

It is difficult to compare results from different methods over different datasets. But to give an impression, current state-of-the-art multimodal approaches are performing around 90% accuracy on the given datasets (Nakamura et al. 2020).

One open challenge is to transfer knowledge from one social media platform, such as Facebook, Twitter, or Reddit, to another one. Each one collects different kinds of data. Not only textual or visual information are important but also meta information is crucial. But exactly this kind of information varies a lot in between the social media networks. The challenge here is to find features that are interchangeable between these platforms, in order to make a method capable to be used in various domains.

Within the previously presented state-of-the-art, many different approaches for tackling the problem of information disorder were introduced and discussed. Many of the proposed methods developed specialized methods, for example Gated-Recurrent-Units (GRUs) (Cho et al. 2014; Dong et al. 2018; Ma, Gao, and Wong 2018) for handling sub-tasks of, or the whole information disorder detection task. This thesis will take well working, established methods, introduced in the following chapter 3, to fuse each modality afterwards of each separate modality, to preserve the informative value and build a robust information disorder classifier.

# 3 A Multimodal Approach for Identification of Information Disorder

This chapter presents all technologies that are going to be used in the experiments. Furthermore, also details about the implementation and evaluation process will be provided. Moreover the theoretical technical background of the used components and model parameters will be introduced and explained in the rest of this chapter.

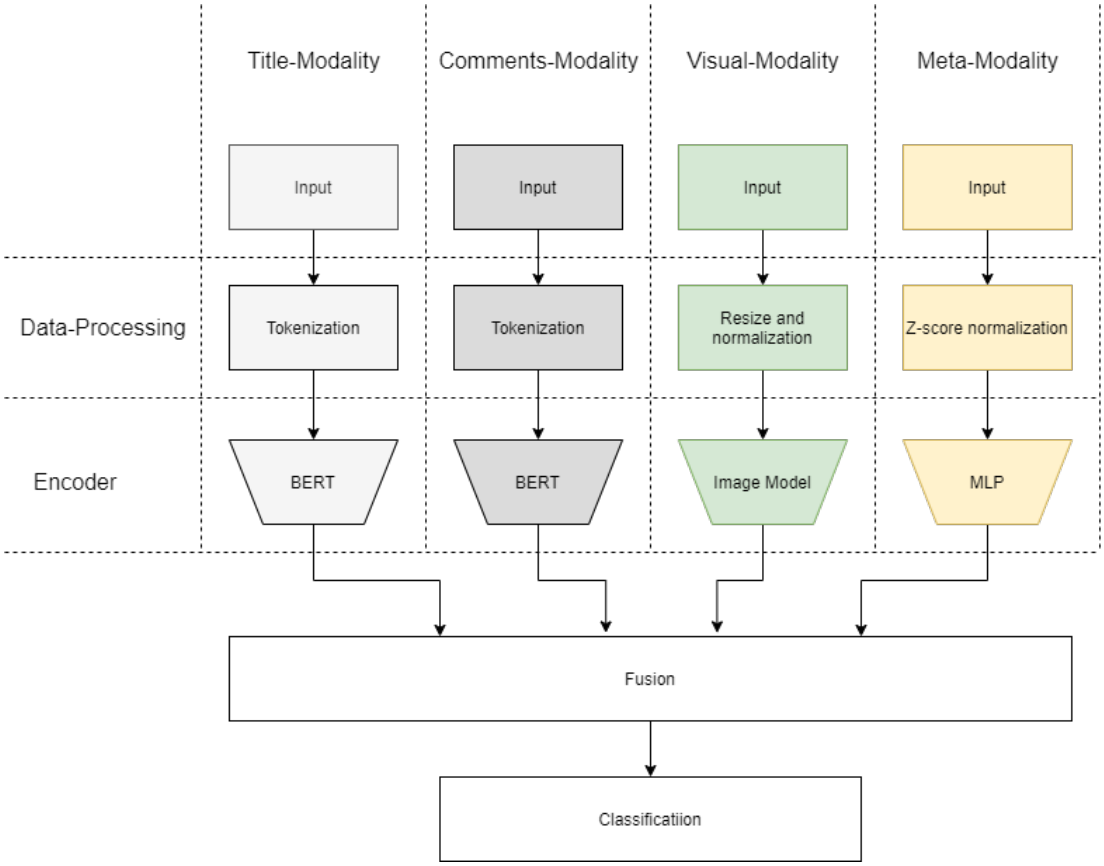


*Figure 3.1.* General method overview of this thesis.

This thesis, as seen in figure 3.1, proposes a four-modality automatic information disorder detection system. The first stage, after reading the data is the data pre-processing step. Depending on each modality different pre-processing steps, such as tokenization or normalization are performed. Afterwards each modality is processed within a separate en-

coder, such as BERT, ResNet50v2, or a MLP. Depending on the experimental setup one or more modalities are fused inside the Level 1 and Level 2 Fusion layer. Afterwards the result is classified into "fake" or "non-fake". An in-depth description of each part and a more detailed description of the method will be given during this chapter.

## 3.1 Input Modalities

This thesis will focus on data that comes from the social media website Reddit. Reddit is a social media platform which allows user to post content concerning different topics. Other user can comment, up- and down vote posts which has impact on the order on which position the post is placed on Reddit. The dataset Fakeddit (Nakamura et al. 2019; Nakamura et al. 2020), which will be introduced in section 4.1, was chosen due its large amount of available multimodal data in the field of information disorder detection.

As shown in figure 3.2 a typical subreddit[1] (1) consists out of $n$ topics (2), which content is textual and/or visual information such as images or videos (2, 3). (4) shows the current upvote/downvote counter and (5) describes the count of related comments for this particular topic.

To summon up, four different modalities can be distinguished. Two textual modalities, which have different semantic meanings, one visual modality and one meta-data modality, which consists out of different post or user describing data. All the before mentioned modalities are part of the chosen Fakeddit dataset, that is going to be presented, with an in-depth analysis, in section 4.1. An example is described below:

As seen in figure 3.3 a tree is shown. The related title (body of the Reddit) is "This tree is split right open" and exemplary one related comment is "exactly what i thought". The meta-data shows 0.94 for the upvote/downvote ratio and 12 for the score.

So each Reddit has one or more modalities:

1. Title of the Subreddits - Text
2. $N$ comments to each related Subreddit - Text
3. Multimedia Content - Image
4. up/down vote ratio, count of comments - Meta-data

A detailed analysis of each encoder for each modality is part of the rest of this chapter.

[1] https://www.reddit.com/r/LanguageTechnology/, retrieved on 17.07.2020.

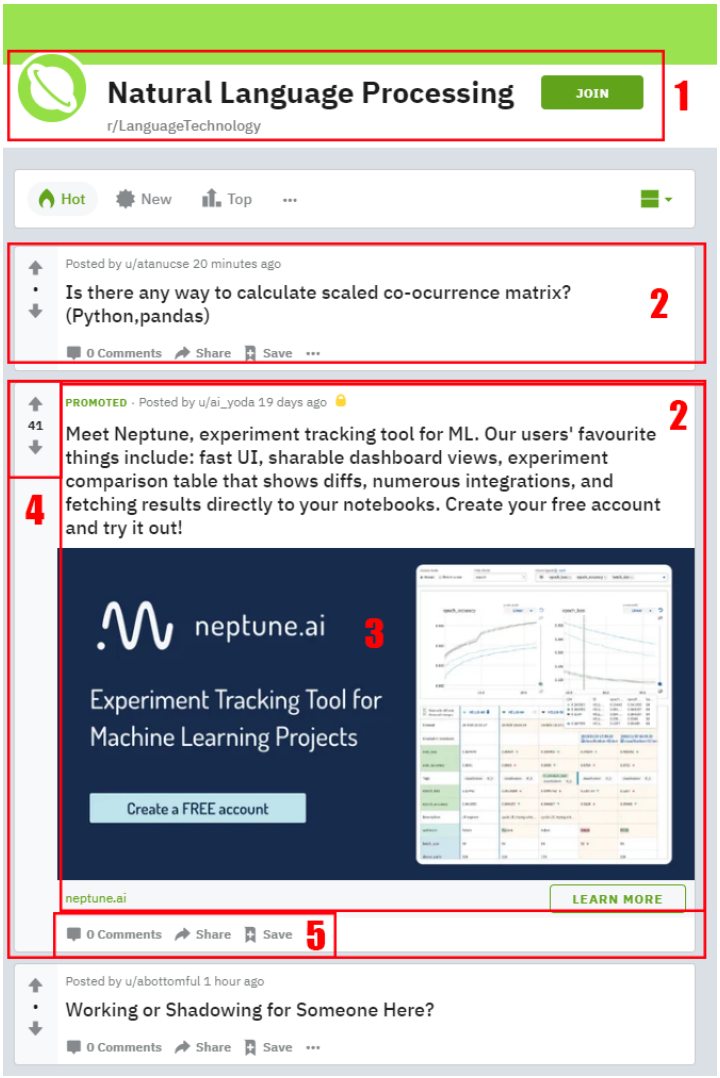


*Figure 3.2.* The structure of a typical Subreddit on the Reddit platform.

*Figure 3.3.* One example of the Fakeddit dataset.

### 3.1.1 Textual modeling

BERT or Bidirectional Encoder Representations from Transformers (Devlin et al. 2018) is a method of learning deep bidirectional representations of unlabeled text. In difference to other state-of-the-art language representation methods such as deep biLMs respectively ElMo (Peters, Neumann, Iyyer, et al. 2018; Peters, Neumann, Zettlemoyer, et al. 2018) or ULMFiT (J. Howard and Ruder 2018) BERT can learn not only from one direction (e.g. from left to right) but also from both direction and benefits from being pretrained on a large corpus of text. Before BERT, NLP methods suffered from on huge disadvantage, the unidirectional learning. This disadvantage should be solved by BERT, which uses a "masked language model" which originates from the 1960s under the name "Cloze procedure" (Taylor 1953). To summarize the idea behind was that the language model must predict the id of randomly masked tokens. BERT extends this method to a so-called "next sentence prediction". This forces the model to learn the connections to both sides of the word respectively sentence. BERT achieves really good results also on new tasks such as Fake News Detection (Nakamura et al. 2020), that is why this method is also chosen for this thesis.

### 3.1.2 Visual modeling

This thesis will perform experiments by using following state-of-the-art image classification networks:[2]

- ResNet50V2
- ResNet101V2
- Inceptionv3

[2] All the networks weights are available via the Keras framework, if desired also pretrained on ImageNet (Deng et al. 2009) https://keras.io/api/applications/, retrieved on 21.07.2020.

### ResNet50v2 and ResNet101V2

The successor of ResNet50 (He et al. 2016a), the ResNet50V2 respectively ResNet101V2 (He et al. 2016b) was chosen for this thesis. The main difference, as seen in figure 3.4 between Version 1 (a) and Version 2 (e) is that in V2 the matrix multiplication with the weights is done after the Batch-Normalisation and ReLu activation. Furthermore, in V2 the last non-linearity is removed. This is a so-called "identity mapping".[3] The difference between the 50 and the 101 version is just the count of residual blocks used within the body of the network.

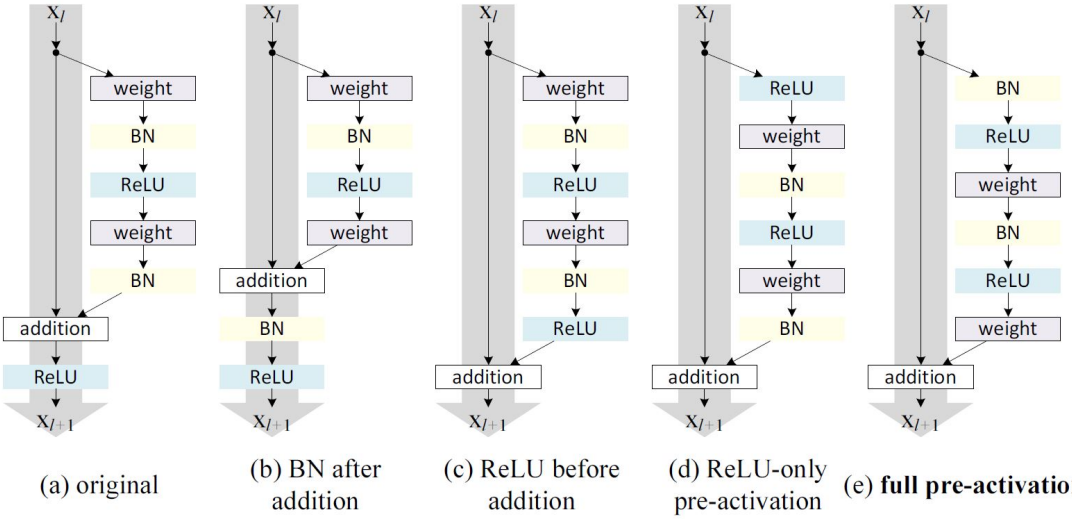


*Figure 3.4.* Difference between ResNetv1 (a) and ResNetv2 (e) with different possible variants in between (b-d), as seen in (He et al. 2016b).

### Inceptionv3

The Inceptionv1 network (former GoogLeNet) was firstly introduced in (Szegedy, W. Liu, et al. 2015). Within the development of the InceptionV2 network (Ioffe and Szegedy 2015) batch normalisation was added. The evolution of the InceptionV3 network was done by adding factorization convolutions with large filter sizes and using auxiliary classifiers (Szegedy, Vanhoucke, et al. 2015). The latest step in the development of the Inception networks was the impact of residual connections on the Inception network. The results was the Inceptionv4 network, the Inception-ResNet-v1 and v2 networks (Szegedy, Ioffe, et al. 2017). Due to its availability within the Keras framework the Inceptionv3 variant was chosen to process the visual modality of this thesis.

[3] For further details see (He et al. 2016b) and https://cv-tricks.com/keras/understand-implement-resnets/, retrieved on 21.07.2020.

### 3.1.3 Modeling of meta-data

Since meta-data can only occur in combination with other data, as they describe them, there are no pretrained or freely available networks like BERT or InceptionV3. In this thesis, a shallow MLP should be established for processing this modality which will be described and analyzed in section 4.1.

## 3.2 Target variable

The dataset provides different kinds of labels, 2-way (fake or no fake), 3-way (true, fake but the text is true) and, 6-way (True, Satire/Parody, Misleading Content, Imposter Content, False Connection, Manipulated Content) (Wardle and Derakhshan 2017) . A detailed analysis of the dataset will be provided in section 4.1. For this thesis the 2-way label is appropriate, but the network will be designed for a multi-label problem.

## 3.3 Multimodal Architecture

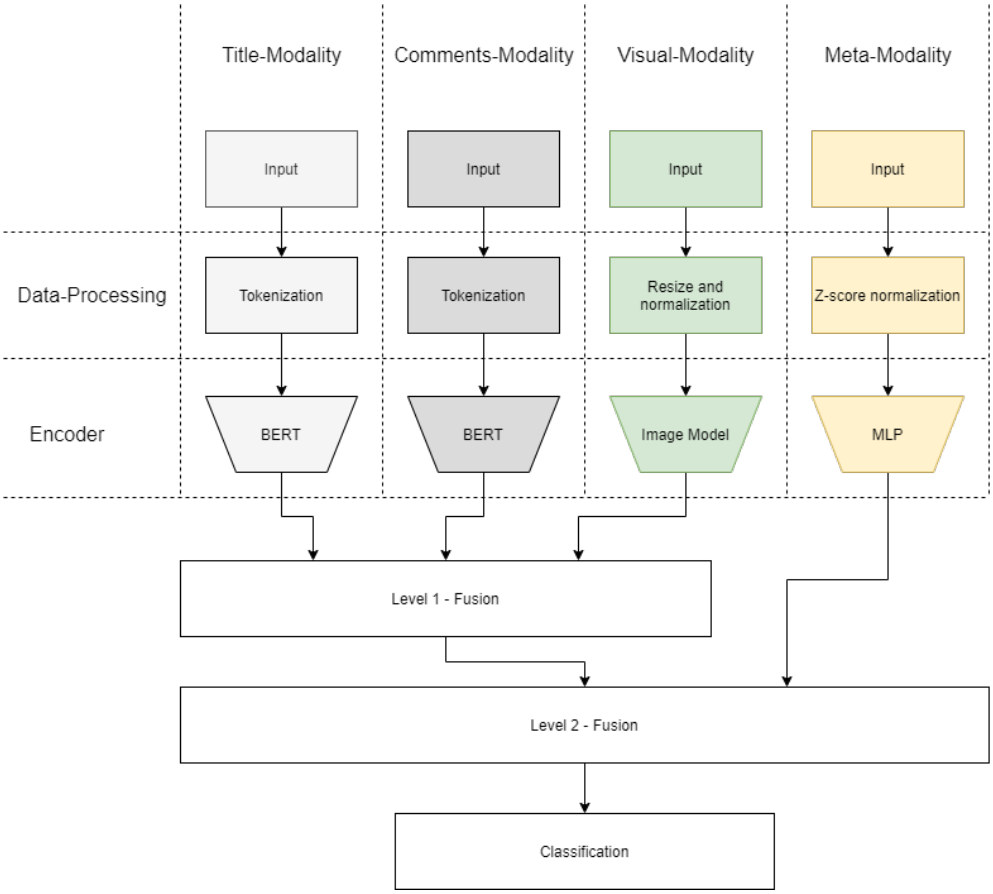


*Figure 3.5.* Detailed overview of the proposed method.

The multimodal architecture is seen in figure 3.5. It consists of several workflows, one for each modality. For the two textual modalities namely title and comments, the data is pre-processed through the BERT Tokenization module. Afterwards the representation is learned by the BERT Language Model. Depending on which modality, or fusion of modalities, is evaluated a fusion layer connects different learned feature vectors. For the visual modality the data is also pre-processed. After resizing and normalising the images, the representation is fed into a state-of-the-art image model, such as ResNet50V2, ResNet101V2 or InceptionV3. Also the meta-data modality is pre-processed by applying z-score normalization to all input vectors. Afterwards the representation is fed to a shallow MLP.

If one modality is evaluated, only one classification layer follows the encoder. If more than one modality needs to be evaluated, a fusion layer is implemented in between. Depending if the meta-data modality is involved, two separate fusion layers (Level 1 - Fusion and Level 2 - Fusion) are implemented to keep the information value of each modality at the same meaning level. The specific implementation details of the fusion layers are described within thr rest of this chapter.

## 3.4 Implementation

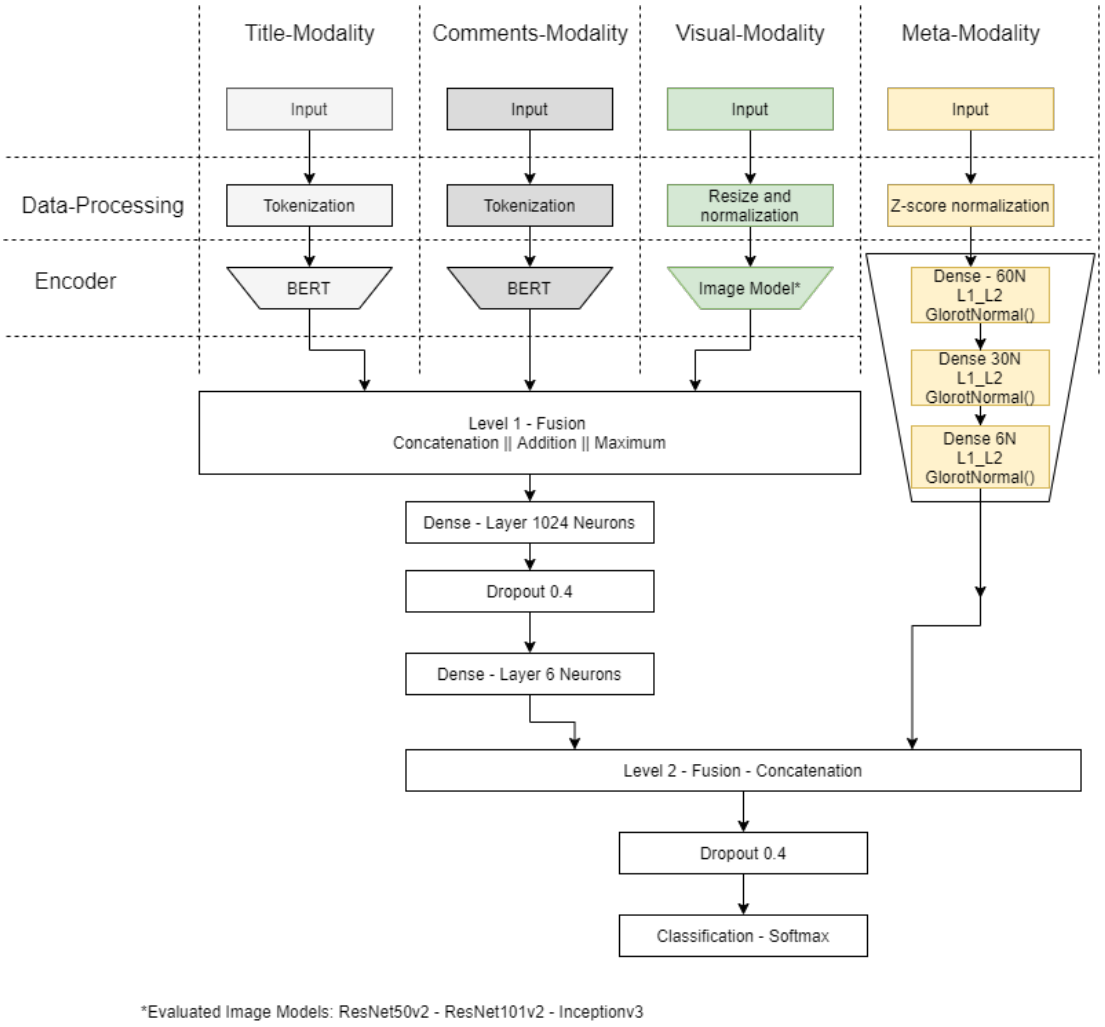


*Figure 3.6.* Detailed overview of the implementation of the proposed method.

The implementation[4] of the thesis was done by using Jupyter Notebooks[5] and Python 3.6.[6] The hardware specifications can be found in section 4.4.2 of this thesis. The neural network was implemented by using the Keras - Platform[7] for Tensorflow 2.[8] For the usage of Tensorflow 2 also the available GPUs should be used addressed by utilizing NVIDIA's CUDA Framework[9] in combination with Tensorflow 2 Mirrored Strategies[10], so the availability of multiple GPUs can be efficiently used. The models itself were implemented by

[4]The repositories for this thesis are for the preprocessing steps: https://github.com/akirchknopf/FID-Preprocessing and for the model / network part: https://github.com/akirchknopf/FID-Model-Handling and for the evaluation part: https://github.com/akirchknopf/FID-Evaluation.
[5]https://jupyter.org/, retrieved on 17.07.2020.
[6]https://www.python.org/downloads/release/python-360/, retrieved on 17.07.2020.
[7]https://keras.io/, retrieved on 17.07.2020.
[8]https://www.tensorflow.org/, retrieved 17.07.2020.
[9]https://developer.nvidia.com/cuda-downloads, retrieved on 17.07.2020.
[10]https://www.tensorflow.org/guide/distributed_training, retrieved on 17.07.2020.

using the Keras Functional API, because it allows building models with several inputs and the ability to fuse different outputs from different models.[11] Unless otherwise stated, RELU is chosen as activation function (Nair and Hinton 2010).

As it can be seen in figure 3.6 the implementation consists out of several components, that are going to be described during the rest of this chapter. For all modalities, except for the meta-data modality, state-of-the-art methods were adapted for processing the data. For simplicity the data-preprocessing is excluded from the follwing network schematics.

### 3.4.1 Data-Processing

Each data modality is pre-processed before being fed into the connected encoder. The textual modalities are pre-processed by applying standard Natural Language Processing methods, such as lower-case all the text and removing punctuation and numbers as described in (Nakamura et al. 2020). The images are resized to fit the neural networks input size and are standardized by the dataset means. The meta-data values were standardized using the z-score normalization approach.

### 3.4.2 Textual component

BERT, as described in section 3.1.1 was implemented by utilizing the "BERT-for-TF2" package.[12] This allows the usage of BERT in form of a common Keras layer. In this thesis the BERT Layer is implemented, followed by a Dropout layer (Srivastava et al. 2014). The classification layer is a classic softmax layer (Bridle 1990a; Bridle 1990b) as seen in figure 3.7.

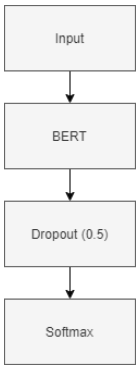


*Figure 3.7.* Architecture overview about the network for the text modalities title and comments.

[11] https://keras.io/guides/functional_api/, retrieved on 17.07.2020.
[12] https://github.com/kpe/bert-for-tf2, retrieved on 17.07.2020.

### 3.4.3 Visual component

All visual classification networks as described in section 3.1.2 and seen in figure 3.8, were implemented in Keras, the weights were pretrained on Imagenet (Deng et al. 2009). During the first part of the experiments all of the three proposed networks are going to be evaluated.

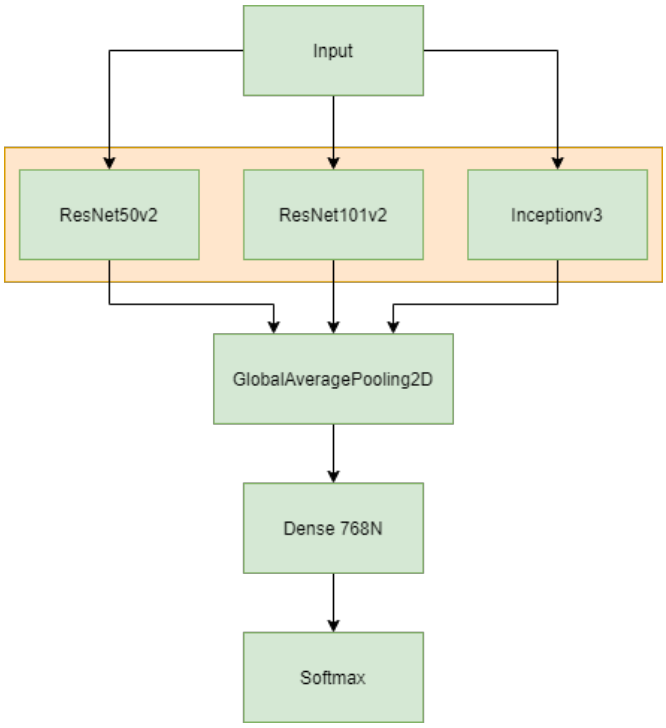


*Figure 3.8.* Architecture overview about the network for the visual modality. For each run a different network ResNet50v2, ResNet101V2 or InceptionV3 is used.

### 3.4.4 Meta component

For the meta component, a small MLP with three layers was created as seen in figure 3.9. The weights were initialized by using the Xavier initializer (Glorot and Bengio 2010). This method successes over "traditional" random initializers. One benefit above the randomly initialized function is the variance of the backpropagated gradient. By using the random initialised weights the variance of the backpropagated gradient is much higher than the variance of the gradient using the Glorot Normal or Xavier initialisation method. On the one hand, this initialising method is very suitable for tanh and sigmoid activation function but, on the other hand not so good for ReLus (Nair and Hinton 2010). Nevertheless, all layers in this thesis are implemented by using ReLu activation. So the meta-model should also being implemented by using this kind of activation function. Furthermore for this shallow MLP L1 and L2 regularizer were also utilized.[13]

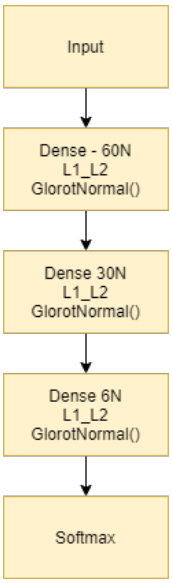


*Figure 3.9.* Architecture overview about the network for the meta modality.

[13] https://keras.io/api/layers/regularizers/, retrieved on 17.07.2020.

### 3.4.5 Two modalities

For the two modalities approach six possible combinations were evaluated. For this purpose, the already trained single modal network has been split up and rebuild. On top of both models, a concatenation layer is set. Afterwards a dense layer with 512 neurons, followed by a dropout layer, and finally a softmax classification layer is build, as seen for the approaches in figures 3.10, 3.11 and 3.12. If the meta modality is involved it is not concatenated to the other modalities directly. Due to its simplicity and availability, this modality will be concatenated after reducing the other modalities ( or the concatenation of it ) to a six neurons representation as seen in figures 3.13, 3.14 and 3.15.

If an image model is evaluated in combination with the other modality, only the best model of the mono-modal runs were further processed.

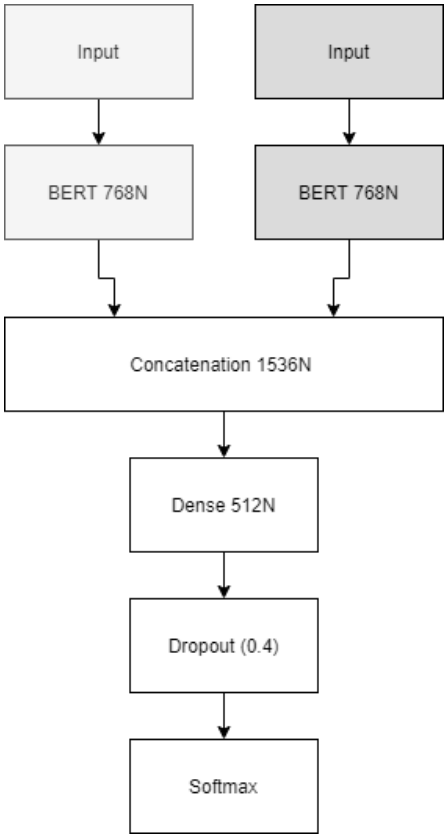


*Figure 3.10.* Architecture overview about the network for both textual modalities.

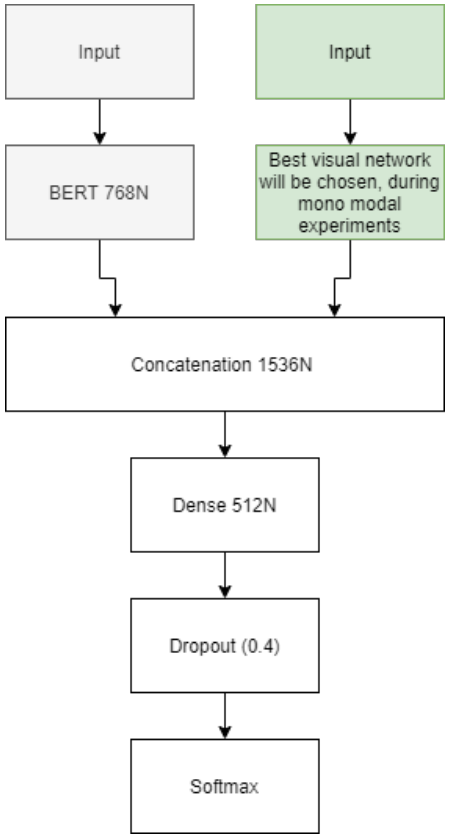


*Figure 3.11.* Architecture overview about the network for the title - visual modalities.

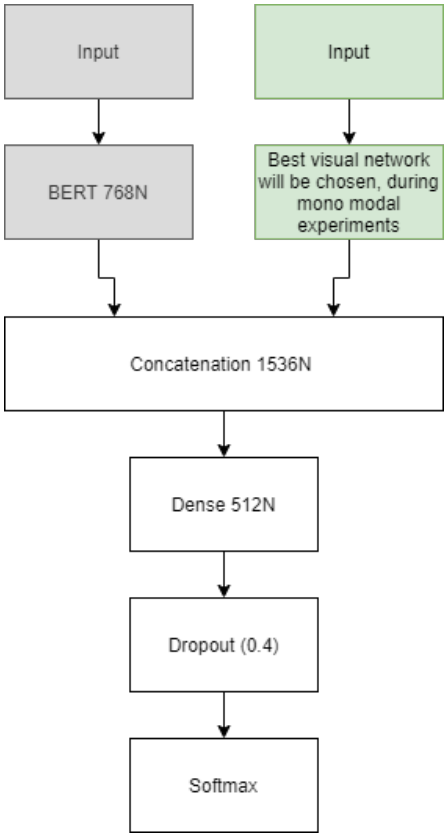


*Figure 3.12.* Architecture overview about the network for the modalities visual and comments.

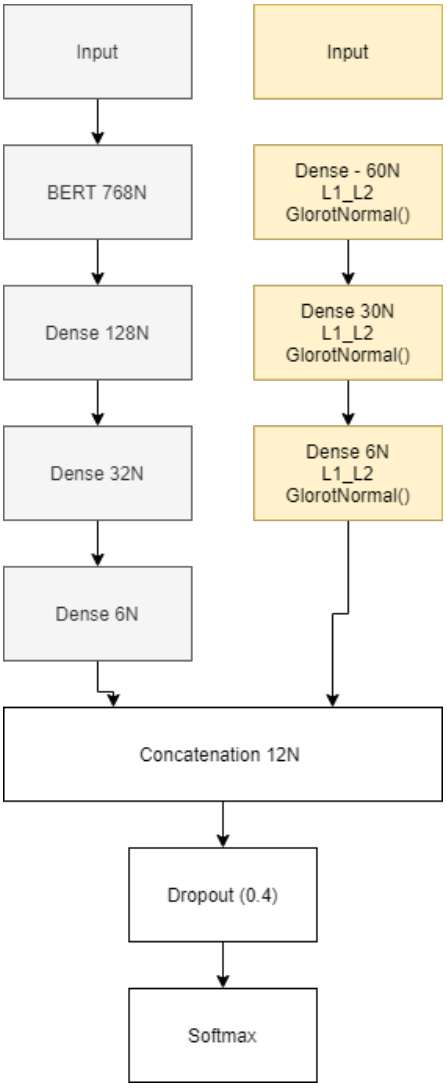


*Figure 3.13.* Architecture overview about the network for the modalities title and meta.

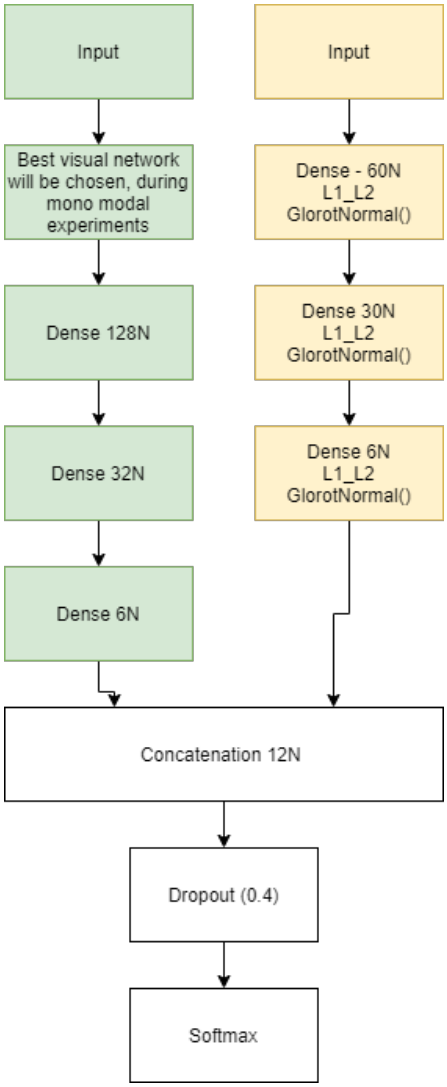


*Figure 3.14.* Architecture overview about the network for the modalities visual and meta.

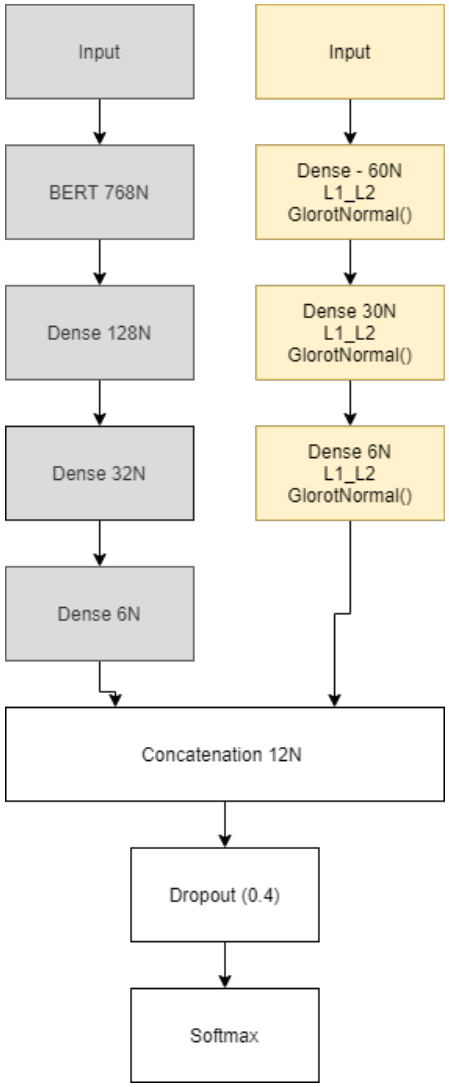


*Figure 3.15.* Architecture overview about the network for the modalities comments and meta.

### 3.4.6 Three modalities

For the three modalities approach four different combinations were evaluated as seen in figure 3.16, 3.17, 3.18 and 3.19. The first architecture "title, comments, and visual" will be also evaluated with different fusion techniques.

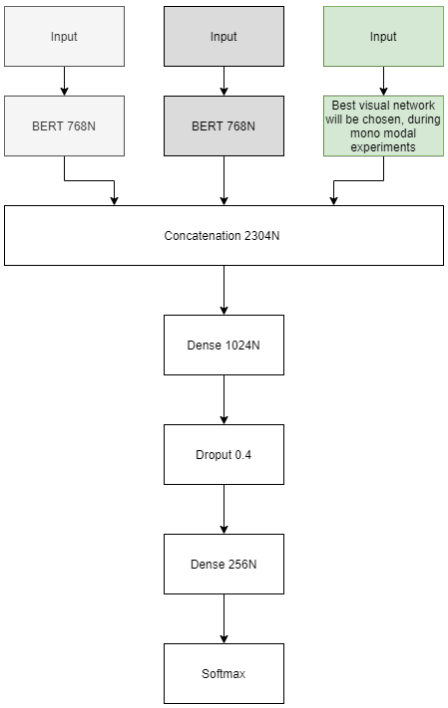


*Figure 3.16.* Architecture overview about the network for the modalities title, visual and comments.

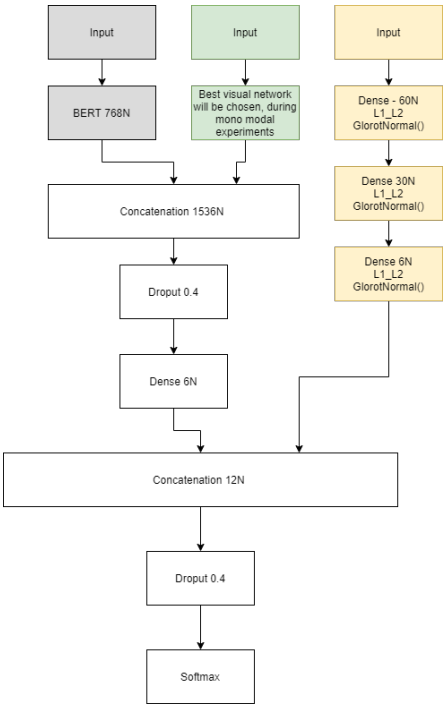


*Figure 3.17.* Architecture overview about the network for the modalities visual, comments and meta.

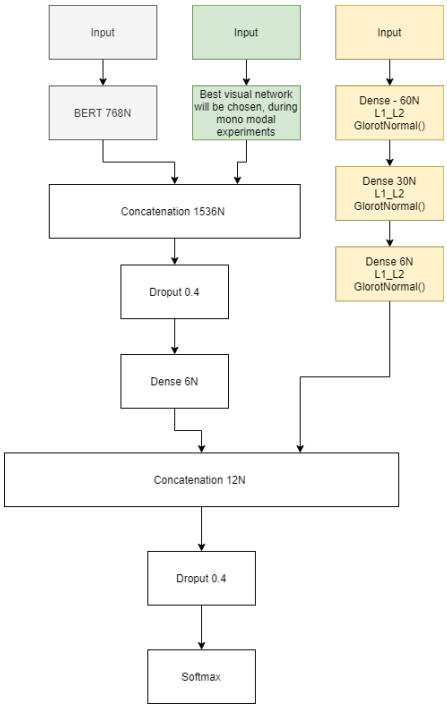


*Figure 3.18.* Architecture overview about the network for the modalities title, visual and meta.

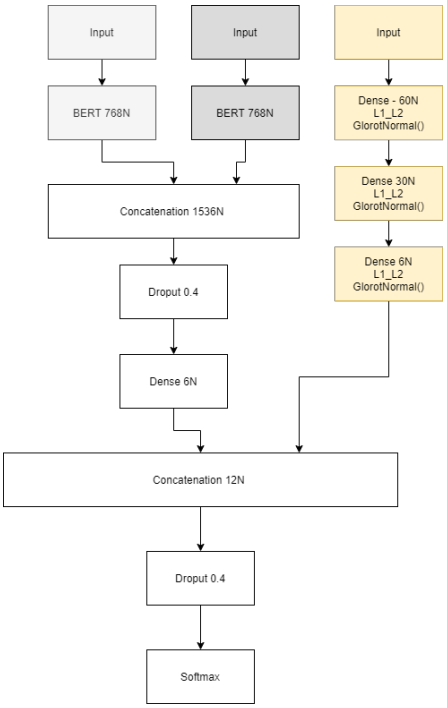


*Figure 3.19.* Architecture overview about the network for the modalities title, comments and meta.

### 3.4.7 Four modalities

For the four modalities approach, all the available modalities are fused and evaluated. The meta modality is again fused in a later step due to its simplicity and shallowness. As seen in figure 3.20 the fusion is again a two level approach as already introduced before to retain the information value of the meta modality.

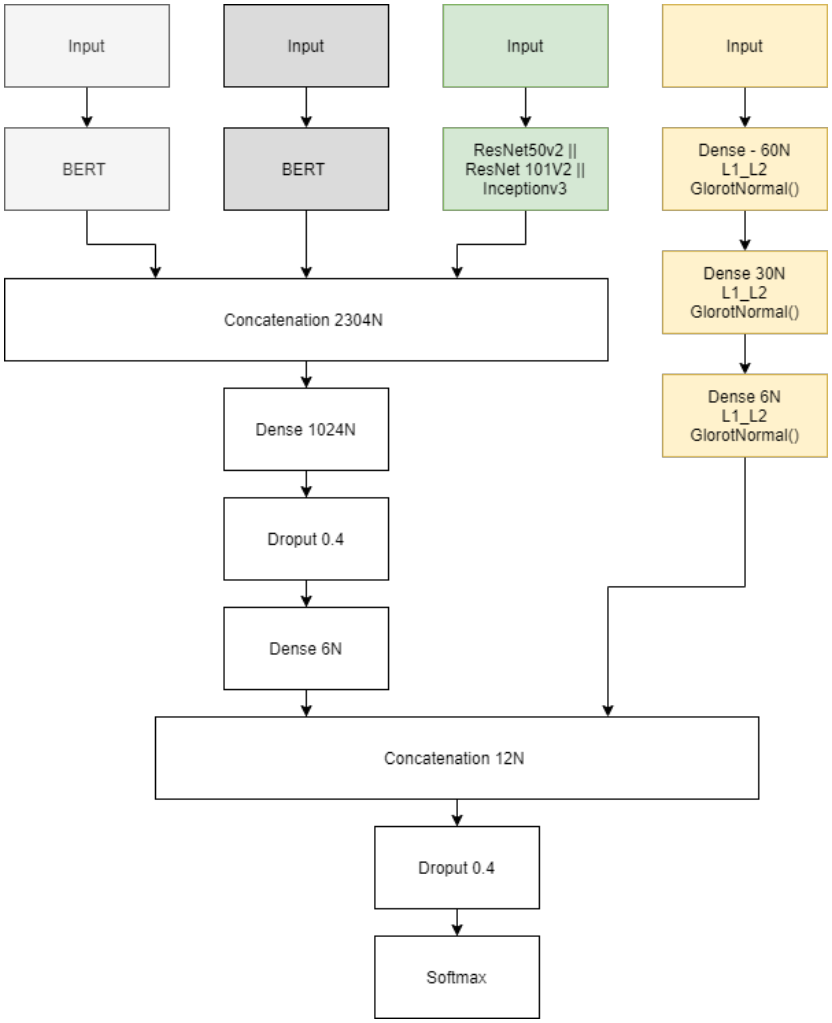


*Figure 3.20.* Architecture overview about the network for the modalities visual, comments, visual and meta.

Within the next sections relevant parts of proposed method such as fusion strategies, loss function and regularization strategies are introduced.

### 3.4.8 Fusion

Within the multimodal approaches, different fusion strategies will be evaluated. Possible candidates are Concatenation but also Maximum or Add fusion strategy would be an option. Only the best multimodal combinations will be evaluated with different fusion strategies which are easily implemented over the Keras API[14]. These are:

- Concatenation
- Maximum
- Addition

#### Concatenation

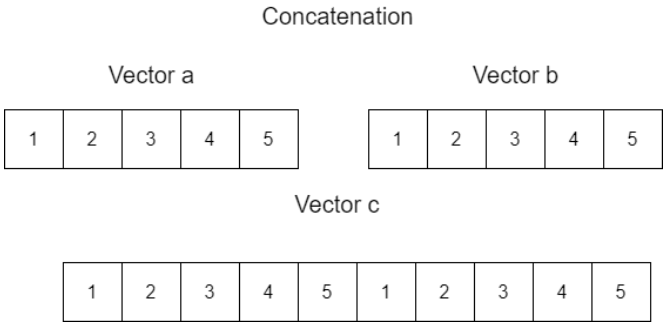


*Figure 3.21.* Example how the Concatenation layer works.

As an example, given is a vector a and a vector b as shown in figure 3.21. Concatenating the two vectors means that the two input vectors, which do not have to have the same length, except for the concatenation axis, are joined lengthwise into vector c, but their values are not changed. This is a common way to retain all information from the layers, or in this case modalities.

[14] https://keras.io/api/layers/merging_layers/, retrieved on 25.08.2020.

### Maximum

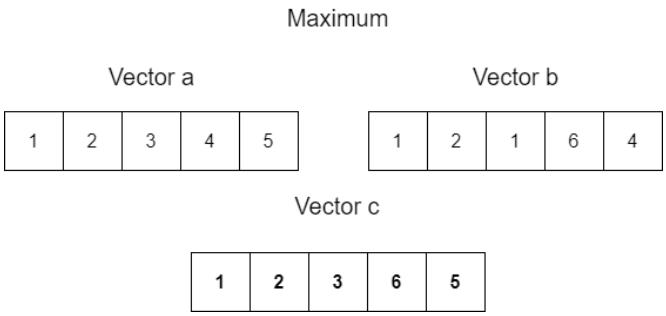


*Figure 3.22.* Example how the Maximum layer works.

As an example, given is a vector a and a vector b as shown in figure 3.22. Maximum of the two vectors means, that both input vectors must have the same shape and after that they are compared element by element and the higher value is copied into a new vector c, with the same shape.

### Addition

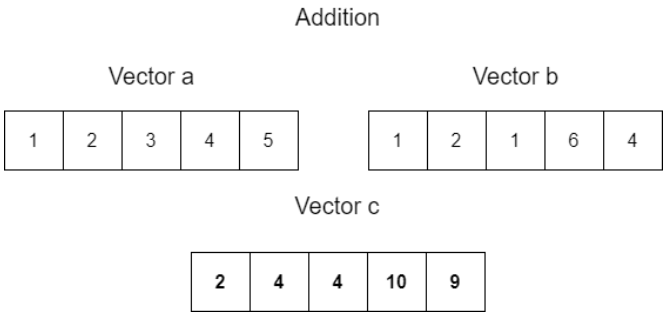


*Figure 3.23.* Example how the Addition layer works.

As an example, given is a vector a and a vector b as shown in figure 3.23. Addition of the two vectors means, that the two input vectors, of same shape, are added element by element and a new vector c. of same shape, is returned.

To summon up, each fusion strategy has it own advantages, which are dependent on the input modality. During the experiments in chapter 4 all of them are evaluated for several models. Only the Level 1 Fusion, as seen in figure 3.6 will be affected by these different strategies. The Level 2 Fusion will be always a concatenation.

### 3.4.9 Loss function and optimization

All the used models in this thesis are going to be optimized on the validation loss. The Sparse Categorical Cross Entropy function was chosen as a loss function. It is a softmax activation plus a cross-entropy loss. This loss is used for a multi-class problem. Since the dataset, which will be introduced in section 4.1, has up to six different classes, this loss function suits very well. The first part, the softmax function is defined in Goodfellow et al. (2016) as followed:

$$softmax(x)_i = \frac{exp(x_i)}{\sum_{j=1}^{n} exp(x_j)}$$

This activation function is widely used in uncountable classification problems and is part of multiple deep learning frameworks such as Caffe, PyTorch, and Tensorflow[15].

The categorical cross-entropy loss must be calculated for each class label per prediction and sum up the result, so the formula is:

$$CE = -\frac{1}{n} \sum_{i \in currBatch} (t_i ln y_i + (1 - y_i) ln(1 - y_i))$$

So *n* determines all samples in the batch, *i* iterates over the batch targets *t* and outputs *y* (Skansi 2018). The difference between categorical and sparse categorical cross-entropy loss is only the format of the labels e.g.:

```
Categorical Cross Entropy:

y_true = [[0, 1], [1, 0], [1,0]]

will changed in Sparse Categorical Cross Entropy to:

y_true = [0, 1, 1]
```

So if the labels are available in integers ( e.g. in the 6-way label ) then sparse categorical cross-entropy is the better choice due to the fact that the conversion from integer label encoding to one-hot encoding have not to be done. The loss function is implemented using the Keras API.

[15] https://keras.io/api/layers/activations/, retrieved on 29.07.2020.

### 3.4.10 Regularizing strategies

In this section strategies for the regularization of neural networks, which are going to be used in this thesis, will be introduced.

#### Dropout

Srivastava et al. (2014) introduced the concept of a Dropout Layer which helps to prevent neural networks from overfitting. The idea behind was pretty simple but very effective. During training the Dropout Layer drops randomly (based on a passed value) unit and their connections to the other neurons. So the networks have to learn in different ways by using alternative connections each time the neurons got deactivated as seen in figure 3.24.

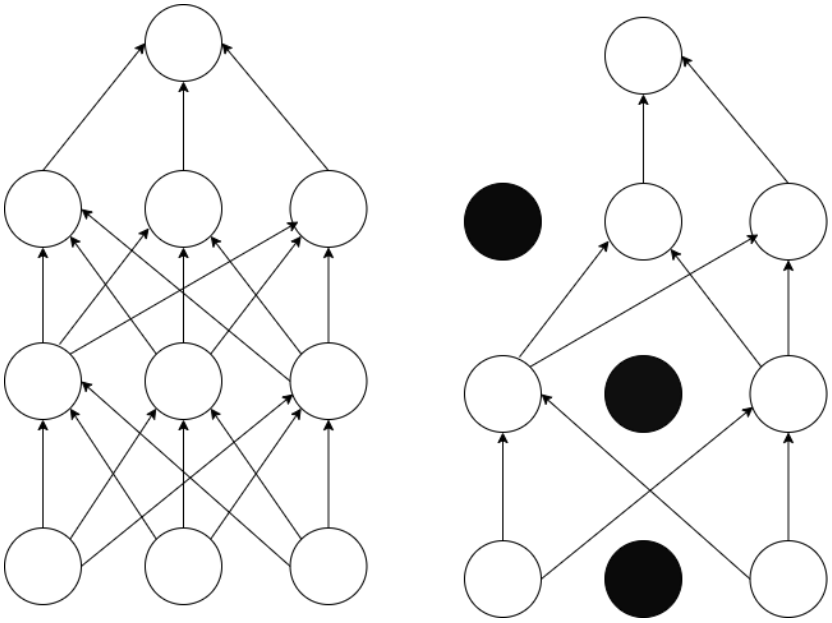

*Figure 3.24.* Example of how Dropout works.

### L2 - L1 Regularization

Originally known as weight decay, ridge regression, or Tikhonov regularization is the L2 regularization one of the most commonly used regularization strategies. The formula is:

$$\Omega(\Theta) = \frac{1}{2} \|\omega\| \frac{2}{2}$$

The L1 regularization is another method of penalizing the model parameters. The formula is:

$$\Omega(\Theta) = \|\omega\|_1 = \sum_i |w_i|,$$

The values of L2 and/or L1 are added to the loss term to prevent the model from overfitting too quick. (Goodfellow et al. 2016) Within the Keras framework regularizers can be easily used by calling the regularizers class/function.[16]

Within this thesis, the combination of L1 and L2 regularizers is used for the meta modality because there is no pretrained network for this modality available and the network should not overfit too quickly.

[16] https://keras.io/api/layers/regularizers/, retrieved on 21.07.2020.

# 4 Experiments & Results

This chapter introduces the chosen dataset Fakeddit and related data sanitation and data pre-processing steps. Afterwards the experimental setups of each modality and combination of modalities are presented including the calculated results.

## 4.1 Dataset: Fakeddit

### 4.1.1 Description

The Fakeddit dataset (Nakamura et al. 2019; Nakamura et al. 2020) was chosen because it was published very recently and consists out of a remarkable amount of multimodal data. This allows experiments on a large scale problem. All types of information disorder, as defined in section 2.1, are part of this multimodal dataset. Another advantage is that due to the novelty of the dataset, there are few methods from other authors available.

The Fakeddit dataset (Nakamura et al. 2020) consists out of one million samples from up to six different categories of information disorder and was collected by using the pushshift API. [1]

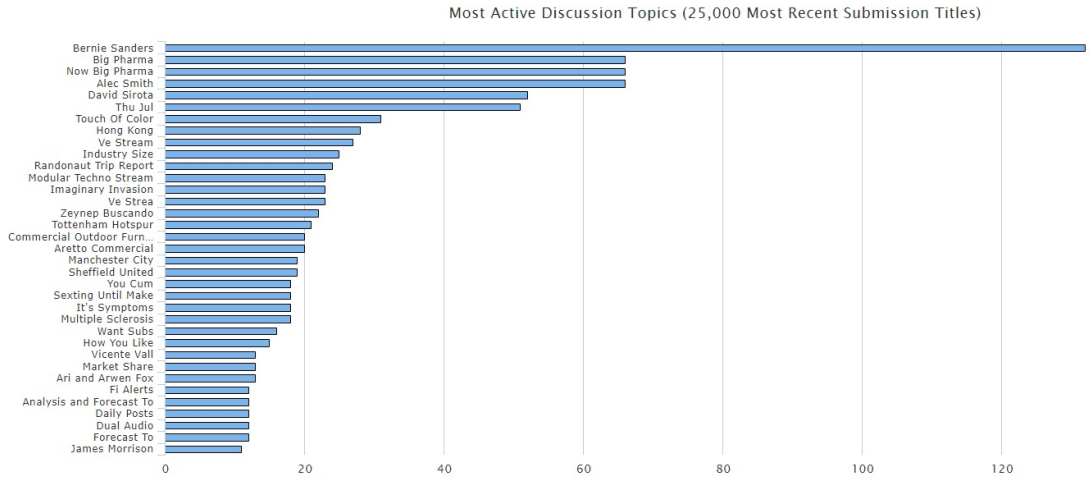


*Figure 4.1.* Excerpt of top 25 most discussed topics on Reddit on 02.07.2020, 13:55.

[1] https://www.pushshift.io, retrieved on 29.07.2020.

Included are not only textual information but also visual and meta-data information.[2] All this information are collected from 2008 to 2019 and around 680k samples are multimodal samples. The source of this dataset is Reddit[3], the so-called front page of the internet. It consists out of uncountable subreddits, or sub-themes where users can discuss and comment on a wide variety of content. An example of what is currently (02-07-2020, 13:55) mostly discussed on Reddit can be found in figure 4.1 but it changes within minutes. It is a very good example of how volatile the current important topics on the internet are. The dataset content is widely spread from political news to daily-life content. Every user can up/down vote and comment each single submission and helps so to separate valuable content from fakes or unnecessary information.

Reddit, as a social media website, has 430 million users[4] and had 1.55 billions total visits from February 2020 to July 2020.[5] With around 130k active communities (subreddits) and around 30 billion page views per month Reddit has a highly active community, discussing about nearly anything and anyone. Depending on the time of the day, around 400k comments and around 50k new submissions are posted on Reddit, as seen in figure 4.2.[6] According to statista.com, Reddit belongs to the Top 6 of the most popular social networks apps in the US with around 50 million US users in September 2019.[7] In July 2020 around 51% of the users originated from the US, followed by the UK with around 7,7%, the first, mainly German speaking country is Germany with around 3.18% on the 5th place.

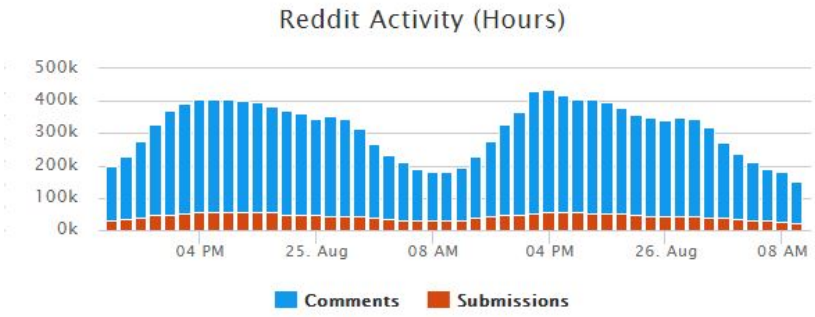


*Figure 4.2.* Overview about the activities per hour of the Reddit community from 25.08.2020 9 am to 26.08.2020 9 am.

[2] The dataset is available here: https://github.com/entitize/Fakeddit.
[3] https://www.reddit.com/, retrieved on 29.07.2020.
[4] Statistics and many more can be found here: https://www.oberlo.com/blog/reddit-statistics, retrieved on 26.08.2020.
[5] https://www.similarweb.com/website/reddit.com/#overview, retrieved on 26.08.2020.
[6] https://pushshift.io/, retrieved on 26.08.2020.
[7] https://www.statista.com/statistics/248074/most-popular-us-social-networking-apps-ranked-by-audience/, retrieved on 26.08.2020.

### 4.1.2 Ground-truth verification

The dataset authors (Nakamura et al. 2019; Nakamura et al. 2020) ensured the data quality by applying a multi step data quality verification process. At first, each moderator of each subreddit checks if all posted items are related to the subreddits topic purpose and removed it otherwise. Secondly, the score feature of each subreddit is used to filter the content furthermore. Every sample with a score below 1 was consequently removed, because the score feature shows if the sample contributes to the topic or is off-topic or misleading.

Due to the large number of samples, the ground-truth verification process is not completely manual due to lack of human resources. The authors took randomly ten samples from each subreddit and checked if the samples are related to the subreddits theme or not. If any of the samples are not related, the whole subreddit is removed together with all related samples. Afterwards all samples are labelled with a 2-way, 3-way, and 6-way label, which will be described in the rest of this section. However, it must be pointed out that not every sample has been manually checked for correctness!

The provided label groups of the dataset are:

**2-way-label:**

- Fake
- No Fake

So the output of a model can be either fake or no fake. This label group does not distinguish if the image is showing a correct content but only the text is false. This will be the preferred label group during the performed experiments in this thesis.

**3-way-label:**

- True
- Fake, but Text is true
- Fake and Text is also a fake

The 3-way label group allows to distinguish results in a more fine-grained way. So the result can be completely true or if the image contains fake information but the describing text can be true. This can happen if the image is manipulated and the text describes the non manipulated image. Finally, the whole sample can be a complete fake, including images and text.

**6-way-label:**

- True
- Satire/Parody
- Misleading Content
- Manipulated Content
- False Connection
- Imposter Content

The 6-way-label group is based on the work of Wardle (2017) as already described in section 2 and shown figure 2.3. Again, there is no hard border where mis-, or disinformation starts or ends, because it is a continuum and sometimes a sample could be interpreted to contain parts of both worlds.

### 4.1.3 Dataset partition

This section describes the technical and statistical aspects of the chosen Fakeddit dataset. In general deep learning datasets are splitted into three different parts. The training set contains most of the samples from the dataset and is used to train the model. The validation set can be either a subset of the training set (which is then not used for training) or a separate set. It is used for the validation step of the model to find out how well the model performs. The third part is the so-called test set. The test set is never touched, except after training the model to test the performance on an unseen part of the data set.

The whole dataset is pre-splitted, by the creators, into four sets: [8]

1. Train set with 878218 samples
2. Validation set with 92444 samples
3. Public test set with 92444 samples
4. Private test set with 92444 samples for the leaderboard (Which is held back) [9]

All three publicly available sets are exactly the same if the three label groups are considered as it can be seen in table 4.1. It should be noted that the test set was also analyzed for completeness only, but no parameters were selected or optimized based on the test set. This would contradict the scientific approach.

[8] These numbers came from the original paper published by (Nakamura et al. 2020).
[9] This is not yet available at the time of writing of this thesis.

| Label-group | Label | Training-Set | Validation-Set | Test-Set |
|---|---|---|---|---|
| 2-way label | Fake | 46% | 46% | 46% |
| | No fake | 54% | 54% | 54% |
| 3-way label | No fake | 46% | 46% | 46% |
| | Fake, but true text | 2% | 2% | 2% |
| | Fake | 52% | 52% | 52% |
| 6-way-label | True | 46% | 46% | 46% |
| | Satire/Parody | 5% | 5% | 5% |
| | Misleading Content | 16% | 16% | 16% |
| | Imposter Content | 3% | 3% | 3% |
| | False Connection | 27% | 27% | 27% |
| | Manipulated Content | 3% | 3% | 3% |

*Table 4.1.* Percentage distribution of the individual label groups on the training, validation and test set

### 4.1.4 Samples

Within the following section some examples are introduced and described. Some of the shown samples are not easily classify for us humans into fake or no-fake and shows how challenging the task in general is.

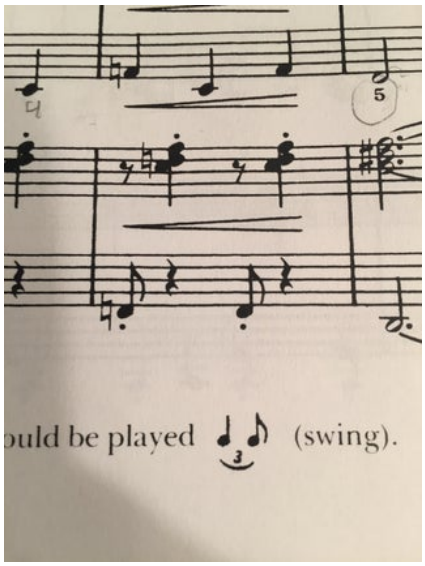


*Figure 4.3.* Sample 1

*Figure 4.4.* Sample 2

As seen in figure 4.3 the image shows a part of a music sheet. The title, "I found a face in my sheet music too", cannot be exposed as fake at first sight. At second glance, however, there is an interesting arrangement of single notes next to the word (sing). If you look closer you can recognize a smiley or a face. The related score is 13, the number of comments is 2, and the upvote/downvote ratio is 0.84.

As seen in figure 4.4 a farm landscape with some sheep is shown. If you take a closer look at the shadow of the sheep on the left side, you will notice that it does not look natural. On closer inspection the shadow turns out to be a badly retouched version of a chicken. The related score is 4, and the number of related comments is 60, and the upvote/downvote ratio is 0.94.

As seen in figure 4.5 several persons in costumes are shown. The title, "The minuteman of Kiev", describes what we can see on the image itself. But such heroes are, if gladly desired, unfortunately fiction and thus fake. This sample does not have a related score, the number of comments is 103, and also the upvote/downvote ratio is not available.[10]

As seen in figure 4.6 a piece of meat is shown. The title, "Rabbi meat from cloned pig

[10] Some meta-data of samples are missing. This is handled during the pre-processing step, as described in section 4.1.6.


*Figure 4.5.* Sample 3

*Figure 4.6.* Sample 4

could be kosher for jews to eat with milk", looks like a fake at first glance. But it is no fake, this image, as also the title were published in an israeli newspaper in 2018.[11] The related score is 577, and the count of comments is 6769, and the upvote/downvote ratio is 0.92.

To summarize this section, it is on the one hand often very obvious to us what is fake and what is not fake, but often enough our eyes are not easily convinced of the truth. The meta modality can support the decision process because there is a clear trend. If you look at the score, you can see that it is increasing and parallel to this the upvote/downvote ratio is also increasing. This indicates that the community has reacted very strongly to this topic, which is especially evident in the last example with almost 7000 comments. It remains to be seen whether this first insight is reflected in the results of the method or not. This is also a challenge for the proposed method.

### 4.1.5 In depth analysis

An in-depth analysis of the dataset showed the following details which will have a great impact on further method development. Many of the best settings for the hyperparameters of the method which should be developed can be estimated in a preprocessing step. The sequence length of the BERT tokenizer is one of the most important hyperparameter for the text and comments modality. The greatest impact has this parameter on the training time and accuracy. If it is too short, valueable parts of the text are cut off, if it is too long the whole training process is stretched unnecessary because of a lower batch size. An analysis of the data showed that 75% of the textual data of the "clean_title" column of the dataset has a length of 57 characters. A visual representation can be found in figure 4.7. The same analysis has been performed on the validation and test set, which shows a very similar picture. It has to be noted that all y-scales are logarithmic scales.

The number of words per title was similarly analyzed. Interestingly, in 75% of cases the titles are not longer than 10 words as seen in figure 4.8. Another analysis was performed on

[11] https://www.timesofisrael.com/rabbi-meat-from-cloned-pig-could-be-eaten-by-jews-with-milk/, retrieved on 26.08.2020.

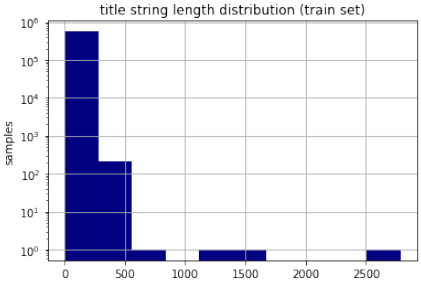


*Figure 4.7.* Distribution of title length of the whole train set.

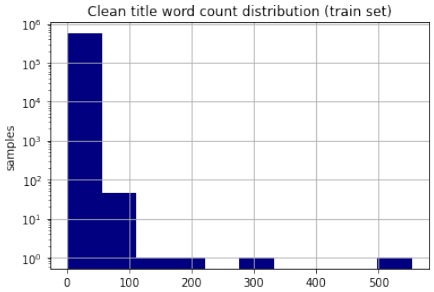


*Figure 4.8.* Distribution of title word count of the whole train set.

the sequence length of the BERT tokenizer. As seen in figure 4.9, 75 % of the sequences are not longer then 138 tokens.

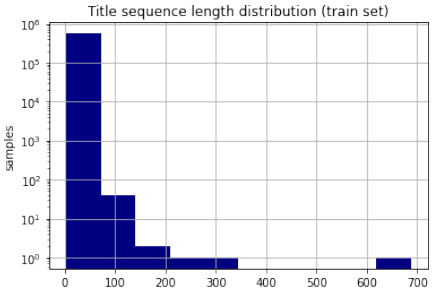


*Figure 4.9.* Distribution of the BERT sequence length of the whole train set.

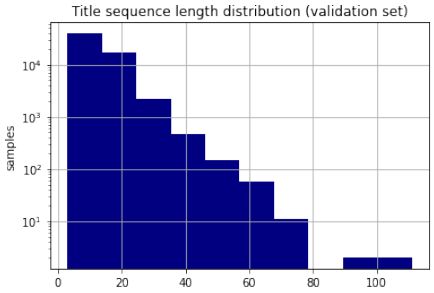


*Figure 4.10.* Distribution of the BERT sequence length of the whole validation set.

A very similar, homogeneous picture regarding the sequence length, can be found in the validation set as seen in figure 4.10. Again around 75% of the tokens were not longer than 140 tokens.

The same analysis was made for the comments of each submission. Around 75% of the comments have a character length of less than 517 chars. A visual representation can be found in 4.11. Another analysis on the word count showed that also 75% of the comments are not longer than 71 words (figure 4.12).

As before the same analysis on the sequences of the BERT tokenizer was performed. As seen in figure 4.13, 4.14 75% of the comments of the train and validation set has a sequence length around 142 tokens.

It must be explicitly mentioned at this point that this analysis was not performed on the test set. For the sake of science, the test set should not be analyzed, since the training parameters should be selected on the basis of the training set and were selected in this thesis.

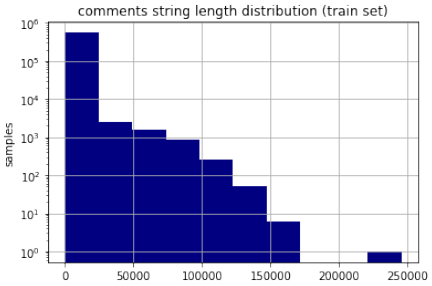


*Figure 4.11.* Distribution of comments length of the whole train set.

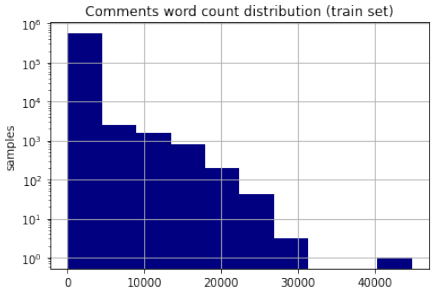


*Figure 4.12.* Distribution of comments word count length of the whole train set.

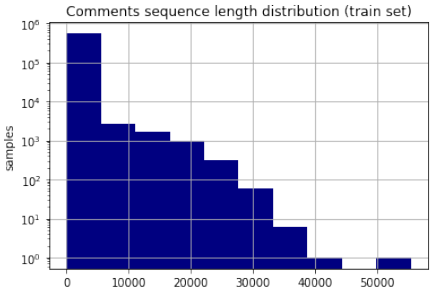


*Figure 4.13.* Distribution of comments sequence length of the whole train set.

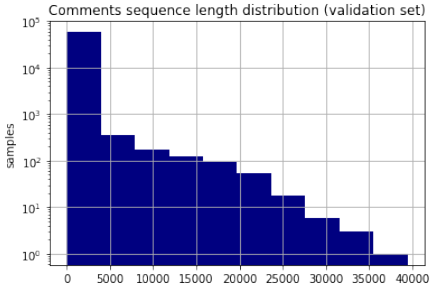


*Figure 4.14.* Distribution of comments sequence length of the whole val set.

### 4.1.6 Data sanitation and pre-processing

In this section the data loading and data sanitation process is described first and then the data pre-processing is discussed. These processes will be described modality by modality. At the end a table shows how many of the samples are left.

#### Textual modality

The data set is made available as a CSV file. The splits are distributed by own csv files except for the submission comments which are collected in one huge 2GB file. This data format is very well compressable but still 800k samples for training are a challenging amount of data. After a first inspection non-relevant columns were dropped. Furthermore, rows with no title attribute are together dropped with samples where no image is available. The creators prepared an already sanitized and converted to a lowercase version of the title within the clean_title column.

For the comments modality, all comments concerning a specific ID are collected, afterwards the text is cleaned as described in (Nakamura et al. 2020). Furthermore, the number of theoretically available and actually existing comments is not negligible. Because these meta information is very important the number of the real available comments is inserted into the column, because the comments text and the number of comments are also processed as a feature.

#### Visual modality

All images could be obtained in two ways. Since all links are available in the csvs files, the images could be downloaded from the respective platforms using a program provided by the authors. Since this possibility often caused problems with non-downloadable images, all images were provided via an archive.[12] The images were available in different image sizes, but all of them were uniformly in JPG format.

Also, the visual information is pre-processed. The images are reduced to a size of 256px x 256px and 768px x 768px. This step is necessary due to the fact that the used image models requires an uniform input image size. This means that images that are too large are scaled down and images that are too small are scaled up. All the images are then saved to a harddisk. During this pre-processing step, the mean and standard deviation is calculated and saved. Afterwards the mean over the whole dataset is calculated and subtracted during runtime together with a normalization between 0 - 1 to be better processed by the neural network. saving the images again has the background, because it would be

[12] Available here: https://github.com/entitize/Fakeddit, retrieved on 27.08.2020.

very expensive to do this on the fly. This would result in an unnecessary extension of the runtime of each epoch. In addition, this process would have to be repeated with each run, which would also be unnecessary.

**Meta modality**

| | author_enc | num_comments | score | upvote/downvote ratio |
|---|---|---|---|---|
| Value ranges | 0 - 273235 | 0 - 10783 | -950 - 137179 | 0.5 - 1.0 |
| No values | 28535 | 167089 | 0 | 167089 |
| % of clean training set | 5% | 30% | 0% | 30% |

*Table 4.2.* Count of samples and percantage of missing meta data entries of the cleaned dataset.

The meta data modality was the most complex one for the data pre-processing step. As seen in table 4.2 an overview about the meta modality of the already cleaned training dataset is given. The first column shows the author encoding. The authors were encoded by replacing the name by an unique id. This resulted in 273234 unique author ids. So in general each author wrote three entries of the cleaned dataset. The last number (273235) is given if the authors name is not available. The number of comments is in the range of 0 to 10783, but for the sake of completeness it should be mentioned that this number refers to the theoretically actual available comments. During the pre-processing step, the actual number of comments available is inserted in this column. This has to be noted also for the score column. The values within this column reaches from -950 to 137179 and shows the immense range of values. The last column, the upvote/downvote ratio ranges from 0.5 to 1.0.

| | author_enc | num_comments | score | upvote/downvote ratio |
|---|---|---|---|---|
| Value ranges | 0 - 273235 | -0.218 - 11.63 | -0.441 - 44.82 | 0.5 - 1.0 |

*Table 4.3.* Meta data values after normalizaton

As already stated above, during this process some problems occurred. Within the columns score, upvote_ratio and num_comments a lot of missing data were observed. Before occurring nan values are filled, the mean and standard deviation were calculated for the z score normalisation. The reason behind is, that a network can learn features which are normalized in the same range or around a certain point more easier. In table 4.3 the value ranges after the normalization are shown. Missing data are filled with so-called NAN[13] values. Afterwards the NAN values, which now encoding missing data, are filled with the negative max values of the column, to represent a identical value that should show anomaly. Additionally a new column is introduced "hasNan". In each row where a NAN value occurs the value is set to 1. The combination of the negative max value in the

[13] NAN = not a number.

original column in combination with the 1 in the additional column should represent the fact, that this data entry was not existent. The goal of this artificial feature generation should be to tell the network in the simplest possible way that all entries with this value combination are "faulty". This is also the reason why the z-normalized columns are not centered around 0, as you can see in table 4.3. This process is also performed for the columns score, upvote_ratio and num_comments.

### Statistics

During the data sanitation and pre-processing step some of the samples were removed from the dataset. Due to the multimodal approach of this method all samples without an image were removed. Sometimes important information such as the title or clean_title were missing. These samples were also removed. These operations were applied to all sets. The following table 4.4 shows the count and percentage of the original number of samples left.

| | Trainings-Set | Validation-Set | Test-Set |
|---|---|---|---|
| before sanitation | 878218 | 76767 | 76752 |
| after sanitation | 560622 | 58972 | 58954 |
| % left: | 64% | 77% | 77% |

*Table 4.4.* Percentage of samples left after the sanitation proces.

To summon up, the data sanitation and pre-processing step is important to normalize and transform the input data into a machine learning understandable data format. The observations and facts established in this section are congruent with those presented by the dataset creators in their paper.

## 4.2 Performance measures

For this method classification accuracy (= portion of correct classified samples) is chosen as a performance measure because for a binary problem, as described in section 3.2, F1 score is possible, but not necessary. Furthermore, the proposed leaderboard[14] and the paper of the dataset authors (Nakamura et al. 2019) and (Nakamura et al. 2020) have also chosen the accuracy metric and thus comparability to alternative approaches is given.

An important factor to mention is the so-called random baseline. The random baseline represents a limit, which means that if this value is exceeded, a learning success is recorded. This limit is 61% in this thesis. In a balanced case this limit lies at 50%. So every result above 61% can be rated as success, everything below shows that the model does not learn from the extracted features.

## 4.3 Hyper-Parameters

This sections introduces the chosen Hyperparameters of the proposed method. Hyperparameters are divided into two different subcatories in this thesis. Firstly, model parameters which controls different settings of parts of the network and related submodels. Secondly, training parameters which controls directly the training procedures.

[14] The leaderboard is at the time of writing this thesis not available, see furthermore: https://github.com/entitize/Fakeddit/issues/1, retrieved on 20.08.2020.

### 4.3.1 Model Parameters

This section introduces relevant model parameters which configures whole modes or only sub settings of the network. These are:

- BERT - Sequence Length
- Image Models - Model
- Image Models - Image Size
- Meta Model - Architecture
- Meta Model - Initializer
- Fusion layer

Already introduces in section 4.1.5 the sequence length is one import model parameter of the BERT model. In the first step, the tokenization, the text is tokenized. The sequence length controls the maximum length of the afterwards taken count of tokens. The maximum value, without changing the BERT layer itself and lossing all the pretrained weights, is 512. The experiments and results in section 4 will show the impact of the sequence length on the models performance.

As already introduced in section 3.1.2 three different image models are going to be evaluated in mono modal runs. The results are shown in section 4.5.1 and the best model will be further used in the multimodal runs. The second related model parameter is the image size. Assuming that a high resolution of the image will result in a high information value and further improve the results. This will be evaluated in section 4.5.1 too.

Concerning the meta model two different model parameters are introduced. At first the whole meta models architecture as described in section 3.4.4. The shallow MLP architecture is chosen due to the available complexity of the input features. Secondly, the initializer function, is chosen based on the paper of Glorot and Bengio (2010) which is also referenced in the Keras Functional API Documentation.

The last model parameter is the fusion layer. Within the evaluation part of this method three different kinds of fusion strategies, namely Concatenation, Addition and Maximum, are going to be evaluated. These are described in detail in section 3.4.8.

### 4.3.2 Training Parameters

For each of the runs, different parameter settings were chosen based on the modality or combination. The following parameters can be varied for training (except for the optimizer, which will be discussed afterwards):

- Batch size
- Epochs of training
- Learning rate

The batch size is being individually adapted for each experiment for fitting on the evaluation hardware which will be introduced in section 4.4.2. This parameter varies from system to system.

The count of epochs for the training will be predefined in the experimental setups during this section and defines how often the the training process will learn from the whole training data. If the early stopping callback triggers before this maximum count of epochs is reached, the experiment will be ended.

Lastly, the learning rate is also a training parameter which can be adapted. A too high learning rate can result in a not converging model and/or in a consequently high loss. Smith (2017) showed that a estimation of the correct learning rate in advance can reduce the training time tremendously. All models in this thesis will be evaluated with a static starting learning rate of $10^{-5}$ for the sake of comparison. If the model is detecting a loss plateau the learning rate is reduced to find the next local minimum.

### Optimizer

As the successor of AdaGrad (Duchi et al. 2011) and RMSProp[15], *Adam* by Kingma and Ba (2017) was chosen due to its simplicity and computationally efficient design as an optimizer algorithm and combining the advantages of both methods mentioned before. The design allows us to calculate the gradient efficiently and is suitable for large scale problems with a lot of parameters to learn.

The Adam optimizer takes three additional arguments, the learning rate, beta_1, beta_2 and epsilon. Experiments by the authors showed that default values of beta_1 = 0.9, beta_2 = 0.999, and epsilon = $10^{-8}$ are suitable for their experiments.[16] and proposes e.g. for an InceptionV3 Network a value of 1.0 or 0.1 for epsilon. This will be evaluated in section 4.5.1. All the proposed values are taken as default inputs for all experiments in this thesis.

### Overfitting prevention

To prevent overfitting two different methods are utilized. First an early stopping callback and secondly a method for reducing the learning rate on a plateau[17]. Both methods are provided by the Keras API.[18] For the experiments, the early stopping patience of three epochs and restoring the weights of the best epoch was most suitable.

[15] Still an unpublished method, background and further information available here: https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf. Retrieved on 02.07.2020.

[16] The Keras documentation for the Adam optimizer states a different default value for epsilon, to precise: $10^{-7}$ https://keras.io/api/optimizers/adam/.

[17] https://keras.io/api/callbacks/reduce_lr_on_plateau/, retrieved on 02.07.2020.

[18] https://keras.io/api/callbacks/early_stopping/, retrieved on 02.07.2020.

## 4.4 Experimental setup

The evaluation section will describe each run configuration with the chosen Hyper-parameters and its result. Within this chapter a detailed description of the how the research questions are going to be answered is given, together with insights about how the experiments are carried out and in which order. Finally, the results are going to be discussed.

### 4.4.1 Research Questions

As in section 1.3 introduced, the research questions should be answered by performing experiments on the Fakeddit dataset.

**Which modality is more meaningful for information news detection, two different textual modalities, the visual modality or the meta modality?**

After developing a multi modal neural network architecture for all modalities, each modality is evaluated one by one. Afterwards the results are evaluated and this question can be answered.

**To what extent can combined multimodal analysis, as opposed to mono-modal, improve the detection of fake information in social media data?**

Afterwards the best configurations and parameters from the first, single modal runs were selected and combined in the developed multimodal approach. Combinations include two, three and all four modalities. All possible combinations of modalities were evaluated to assess their mutual benefit.

**Which network architectures from research are best suited for the multimodal analysis of fake information?**

To answer this research question different modality-specific pre-trained network architectures were integrated into the proposed solution for encoding the respective modalities. The different architectures were systematically evaluated and compared.

### 4.4.2 Evaluation Protocol

Within the rest of this chapter the experiments are carried out and the results are described. Firstly, the experiments for each modality are described. Beginning from the mono-modal experiments, over fusing two, three and lastly all four modalities. For the first mon-omonodal experiments, the architecture as described in section 3.4 and seen in figure 3.6 is splitted up and each modality is evaluated one-by-one. Afterwards the best mono-modal models are fused into a dual modal setting and their output is fused together within the Level 1 Fusion, an example is seen in figure 3.11. If the meta modality is involved, the models output is consequently reduced and then concatenated and evaluated an example of this fusion strategy is seen in figure 3.13. The same procedure is adapted for fusing three different kinds of modalities and afterwards for all four modalities.

Each experiment is systematically carried out by the following schema. Firstly, the appropriate settings of the experiments are logged and saved for later investigation. Secondly, the model is trained on the training set until the maximum count of epochs is reached, or the early stopping criterion is fulfilled. This will prevent the model from overfitting. Thirdly, after each epoch the models performance is evaluated on the validation set, which is held out and is not part of the training set. The models weights are updated. Lastly, after the training procedure has been completed, the models performance is finally evaluated on the separate test set.

In addition to the logging of parameters and results the best weights are also saved for further evaluation and investigation.

#### Hardware setup and implementation

All experiments are carried out on following hardware:

- Operating System: Ubuntu 18.04.4 LTS
- CPU i7-6900K @ 3.2 GHZ - 16 Threads
- 128 GB RAM
- 1 TB SSD Samsung EVO 860
- 4 x GTX 1080 Founders Edition with 8GB RAM

## 4.5 Experiments and results

### 4.5.1 Mono modal experiments and results

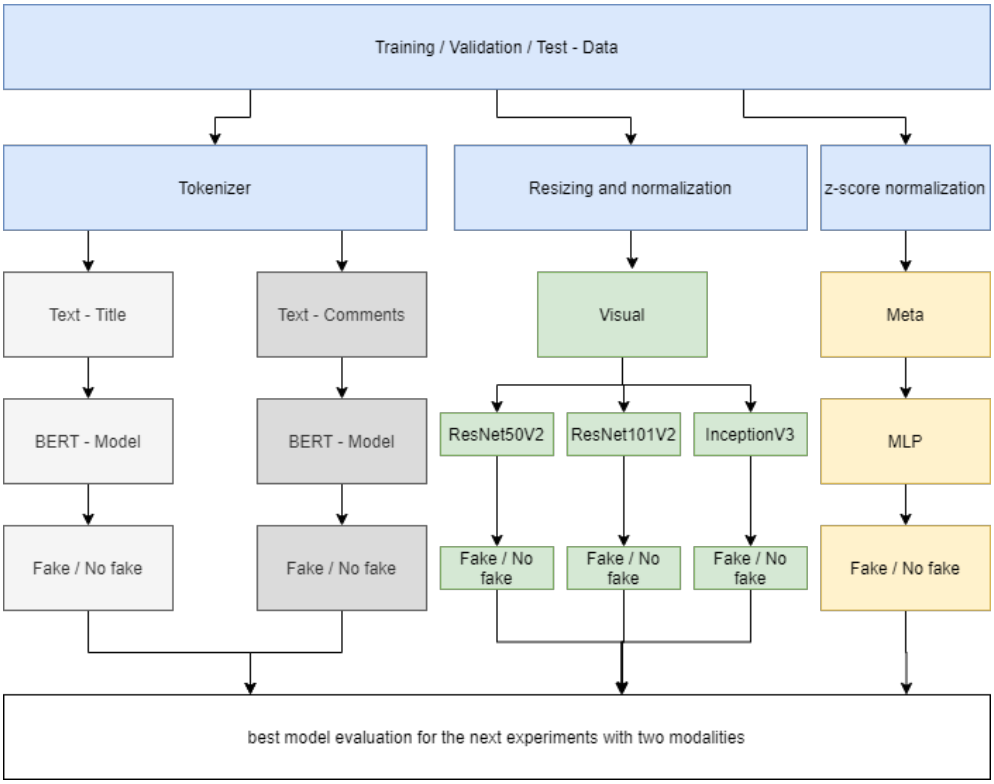


*Figure 4.15.* Overview about the single modal approach

The first experiments will test all single modalities such as Title, Comments, Visual, and Meta as seen in the overview figure 4.15. Afterwards the best parameters and network configurations were chosen for following multimodal experiments. For the combination approaches the pretrained weights of the single modality networks were loaded and only the resulting head is trained due to time and performance reasons. The head describes all layers which are used to fuse and classify the processed samples. It is marked as white boxes with black borders within section 3.4.5 and beyond.

### Title and Comments modality

For the textual modality of the experiments, BERT was chosen as mentioned before. The assumed settings for the textual modality can be found in table 4.5 for the title and 4.6 for the comments. To validate the assumption that the sequence length should be chosen so that most samples[19] can be mapped to it, a simplified form of grid search is used. The aim is to evaluate the model on the given dataset by altering, in this case, the sequence length, to find the best ratio between accuracy and runtime. The results can be found in the following tables.

| | Run 1 | Run 2 | Run 3 | Run 4 | Run 5 | Run 6 |
|---|---|---|---|---|---|---|
| Sequence length | 32 | 64 | 128 | 192 | 256 | 512 |
| Epochs | 7/10 | 7/10 | 5/10 | 5/10 | 5/10 | 5/10 |
| Batch Size | 512 | 228 | 96 | 64 | 32 | 16 |
| Optimizer - Beta 1 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 |
| Optimizer - Beta 2 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 |
| Optimizer - Epsilon | $10^{-8}$ | $10^{-8}$ | $10^{-8}$ | $10^{-8}$ | $10^{-8}$ | $10^{-8}$ |
| Duration (mean) minutes | 15.2 | 32.2 | 66.2 | 101.6 | 158.4 | 352 |
| Accuracy - Validation % | 88.01% | 88.21% | 88.10% | 87.99% | 88.10% | 88.23% |
| Accuracy - Test % | 88.03% | 88.20% | 88.10% | 87.95% | 88.23% | 88.23% |

*Table 4.5.* Run configuration and results for the title modality.

| | Run 1 | Run 2 | Run 3 | Run 4 | Run 5 | Run 6 |
|---|---|---|---|---|---|---|
| Sequence length | 32 | 64 | 128 | 192 | 256 | 512 |
| Epochs | 7/10 | 6/10 | 6/10 | 6/10 | 5/10 | 5/10 |
| Batch Size | 512 | 228 | 96 | 64 | 32 | 16 |
| Optimizer - Beta 1 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 |
| Optimizer - Beta 2 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 |
| Optimizer - Epsilon | $10^{-8}$ | $10^{-8}$ | $10^{-8}$ | $10^{-8}$ | $10^{-8}$ | $10^{-8}$ |
| Duration (mean) minutes | 15.14 | 32.2 | 66.2 | 101.2 | 159.6 | 351.2 |
| Accuracy - Validation | 84.06% | 85.71% | 86.69% | 86.71% | 86.80% | 86.82% |
| Accuracy - Test | 84.10% | 85.63% | 86.52% | 86.67% | 86.70% | 86.75% |

*Table 4.6.* Run configuration and results for the comments modality.

The results showed an interesting picture. For each of the six performed runs for each of the both modalities title and comments a similar behaviour could be observed.

Starting with table 4.5 and figure 4.16, it can be seen that altering the sequence length has impact on various details. The first and most obvious fact is the parallelism between sequence length and duration of the corresponding epoch times. This is, for the title and comments modality nearly the same, and therefore visualized in figure 4.18 and figure 4.19 too.

[19] For the assumption value see section 4.1.

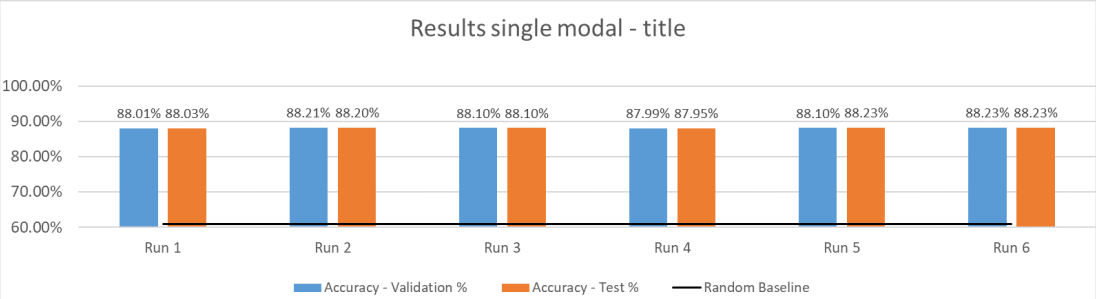


*Figure 4.16.* Results of the single modality title.

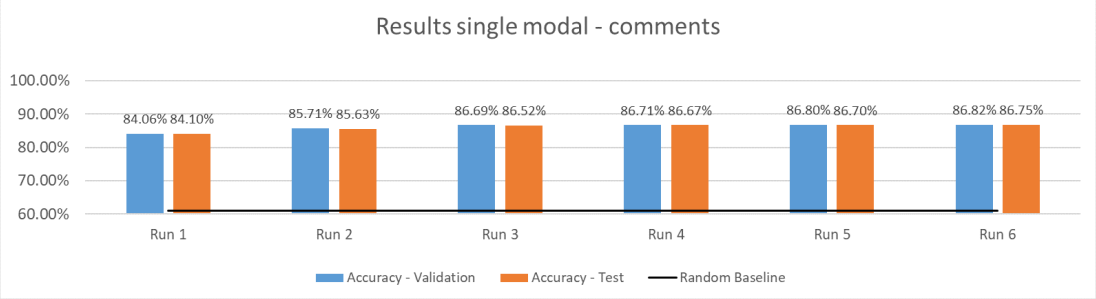


*Figure 4.17.* Results of the single modality comments.

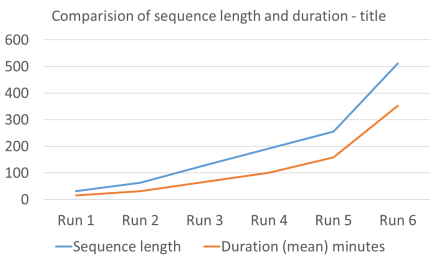


*Figure 4.18.* Results of the evaluation of the ratio sequence length to processing time for the title modality.

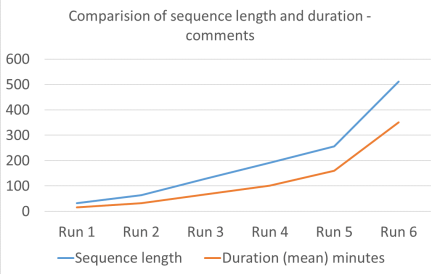


*Figure 4.19.* Results of the evaluation of the ratio sequence length to processing time for the comments modality.

Having a closer look into table 4.5 and compare the sequence length and accuracy on the validation and test set, the accuracies are hardly rising any more. Looking at the results of the comments modality in table 4.6 a similar picture is shown.

The same parallelism between sequence length and epoch duration can be observed as also the moderate increase of the accuracies, which are far over the random baseline, which is at 61%.[20] The results with around 88% for the title modalities and around 85% for the comments modality is very good and shows the strength of the proposed mono-modal approach.

Furthermore, it could be observed that the batch size does not have a lot of impact on the accuracies. That shows that the models have a good ability to generalize and learn from the feature representations. In addition to this, it has to be noted that in no run the maximum count of epochs (10) has been reached and the early stopping callback stopped the training in before.

To summon up, to take a good average between the runs and also considering a moderate training time per epoch, a sequence length of 128 tokens is the best choice for the dataset. Also, the mean of the calculated five epochs with around 66 minutes is moderate and acceptable. This configuration for both modalities are further implemented into the following experiments as model parameters. Furthermore the first observation of section 4.1.5, that the sequence length of BERT should be approximated in a pre-processing step is correct. The results of both textual modalities are very good. The comments modality is even a little better then the title modality, which was surprisingly.

[20] The random baseline is represented in each image by the black line in the lower part of the image.

### Visual Modality

The visual modality will be evaluated by utilizing three different state-of-the-art image classification models as introduced in section 3.1.2, namely ResNet50v2, ResNet101v2 and Inceptionv3. For this purpose different experiment parameters are assumed and introduced in the table 4.7. The main points to evaluate are, next to the different architectures, also the impact of the image size on runtime and accuracy. The detailed results can be found below in table 4.7

| | Run 1 - ResNet50v2 | Run 2 - ResNet101v2 | Run 3 - InceptionV3 | Run 4 - InceptionV3 | Run 5 - ResNet50v2 | Run 6 - InceptionV3 |
|---|---|---|---|---|---|---|
| Epochs | 5/10 | 5/10 | 5/10 | 10/10 | 5/10 | 5/10 |
| Image Size | 256px x 256px | 256px x 256px | 256px x 256px | 256px x 256px | 768px x 768px | 768px x 768px |
| Batch Size | 256 | 128 | 256 | 256 | 32 | 32 |
| Optimizer - Beta 1 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 |
| Optimizer - Beta 2 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 |
| Optimizer - Epsilon | $10^{-8}$ | $10^{-8}$ | $10^{-8}$ | 1 | $10^{-8}$ | $10^{-8}$ |
| Duration (mean) minutes | 35.40 | 56.4 | 47 | 41 | 265.4 | 258.6 |
| Accuracy - Validation | 77.13% | 77.50% | 78.11% | 61.41% | 80.89% | 81.04% |
| Accuracy - Test | 77.47% | 77.62% | 78.48% | 61.43% | 81.23% | 81.51% |

*Table 4.7.* Run configuration and results for the visual modality.

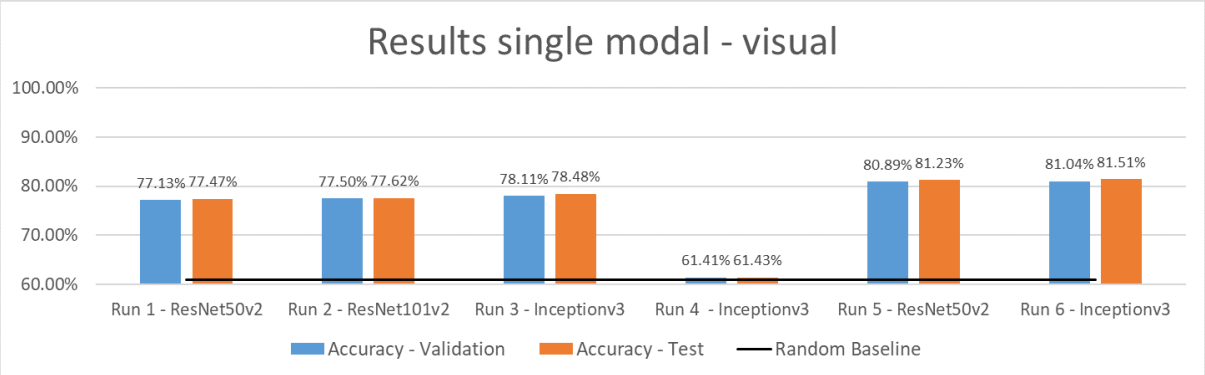


*Figure 4.20.* Results of the single modality visual.

The experiments with the visual modality showed a clear picture. The chosen networks: ResNet50v2 (Run 1 and Run 5) ResNet101v2 (Run 2) and InceptionV3 (Run 3, 4 and 6) performed pretty similar, but the Run 6 with the InceptionV3 was the best. Run 4 shows an outlier, which can be explained, because this run was made under the assumption that a value of epsilon=1 for the optimizer will result in better runs, what was proposed in the Keras documentation, referenced in the footnote of section 4.3.2. The result of 61.4%, which is hardly over the random baseline shows clearly that changing this parameter results in a large drop in the performance of the model. Run 5 and 6 were performed under the assumption that an input image with a higher resolution is favorable for fake classification. This can be confirmed because Run 5 and Run 6 performed on the one hand better with an input size of 768px x 768px (before 256px x 256px) but on the other hand each epoch was around five times slower. From 50 minutes per epoch to around 255 minutes. So a compromise must be made if the better accuracy from around 81% counts or the faster processing time with around 77% accuracy, as it can be seen in figure 4.20. When looking

at the count of trained epochs it can be noticed that again the maximum count, except for Run 4, is not reached. Another option would be to fuse models which are not performing at the best level, but saving a lot of time during the training procedures. In this thesis time is not a problem so the better performing model is taken for all further experiments.

### Meta-Data Modality

The meta modality will be evaluated in a two-step process. The first step will evaluate which of the available features are meaningful, all of them, and then feature by feature. The experiment settings can be found in table 4.8. The second step will fuse the most important features and evaluate if artificially created columns (hasNanScore, hasNanUpvote)[21] have a positive impact on the accuracy of the network. These settings can be found in table 4.8 and table 4.9.

| | Run 1 | Run 2 | Run 3 | Run 4 | Run 5 |
|---|---|---|---|---|---|
| Epochs | 5/100 | 5/100 | 100/100 | 100/100 | 100/100 |
| Choosen columns | all | author_enc | score | upvote_ratio | num_comments |
| Batch Size | 1024 | 1024 | 1024 | 1024 | 1024 |
| Optimizer - Beta 1 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 |
| Optimizer - Beta 2 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 |
| Optimizer - Epsilon | $10^{-8}$ | $10^{-8}$ | $10^{-8}$ | 1 | $10^{-8}$ |
| Duration (mean) minutes | 0.23 | 0.23 | 0.23 | 0.23 | 0.23 |
| Accuracy - Validation | 61.16% | 60.76% | 61.21% | 74.54% | 75.59% |
| Accuracy - Test | 61.19% | 60.41% | 61% | 74.13% | 75.63% |

*Table 4.8.* Run configuration and results for the meta-data modality, only single feature.

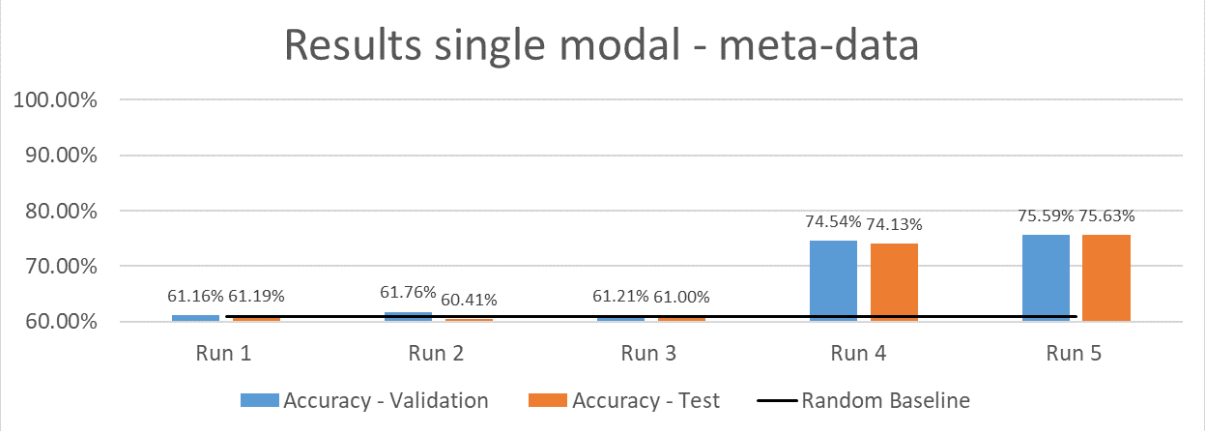


*Figure 4.21.* Results of the single modality meta-data by using in Run 1 all features and in the Runs 2 - 5 only one feature.

The meta modality represents a difficult aspect of the dataset. The data consistence was not very good and there were missing values as described on section 4.1.6. First

[21] Why these columns were created is explained in section 4.1.6.

experiments with all meta-data (Run 1) and just with the author encoding (Run 2) performed not very well, hardly above or below the random baseline. So looking at the results, the author encoding does not provide much information about if the topic is fake or not. The same picture can be observed if all available and pre-processed meta-data are taken into account (Run 1) or just the score feature (Run 3).

The most promising features were the upvote ratio (Run 4) which performed very well with an accuracy on the validation and test set with around 74% and the count of available comments regarding one topic, which performed slightly better with a validation and test accuracy about 75%. The training time was negligible with around 14 seconds with a batch size from about 1024 as it can be seen in table 4.8 respectively in figure 4.21. In this case the count of epochs was always reached. More experiments can show if there is any improvements if the count of epochs are further increased.

| | Run 6 | Run 7 | Run 8 |
|---|---|---|---|
| Epochs | 100/100 | 100/100 | 100/100 |
| Choosen columns | score nan_score | upvote nan_upvote | upvote num_comments |
| Batch Size | 1024 | 1024 | 1024 |
| Optimizer - Beta 1 | 0.9 | 0.9 | 0.9 |
| Optimizer - Beta 2 | 0.999 | 0.999 | 0.999 |
| Optimizer - Epsilon | $10^{-8}$ | $10^{-8}$ | $10^{-8}$ |
| Duration (mean) minutes | 0.23 | 0.23 | 0.23 |
| Accuracy - Validation | 61.21% | 74.54% | 77.8% |
| Accuracy - Test | 60.1% | 74.13% | 77.34% |

*Table 4.9.* Run configuration and results for the meta modality, only best single features.

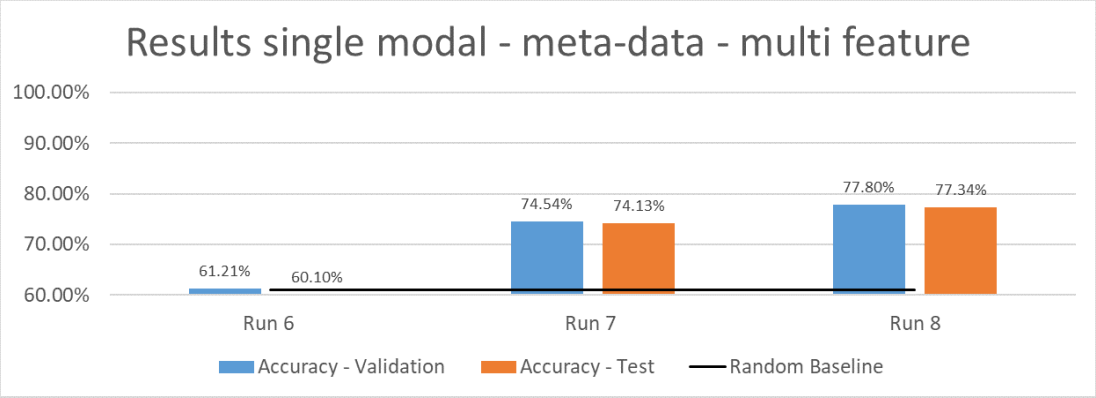


*Figure 4.22.* Results of the single modality meta-data by using the best single features combined.

Fusing the most promising single features, namely upvote ratio and num_comments, together showed an interesting picture. As it can be seen in figure 4.22 in Run 6 combining the features of the score with the hasNanScore feature does not increase the accuracy on the validation and test set. The accuracy is more or less the random baseline with 61%. The same observation could be made considering the upvote together with the

hasNanUpVote feature.The accuracy is more or less the same from the single feature run with around 74%. A different picture could be taken if the most promising features of the single feature meta runs 4 and 5, namely upvote ratio together with the count of comments are combined. After a learning time of around 100 epochs the validation and test accuracy was finally around 77.8% on the validation set and 77.34% on the test set. So combining these feature results in a slightly better run (Run 8) and showed that Reddit's meta data can also provide a useful source for fake information detection. Again the epoch time of around 14 seconds per epoch did not change.

In this section the mono-modal models have been evaluated by choosing different parameters for the training procedures and model parameter for the configuration of the models. All the modalities showed remarkable results on validation and test sets. The two text modalities performed around 88% resp. 86& and showed the importance of textual content in this task. The experiments on the visual modality proved that also images, depending on the image size, can contribute a lot to this task. The best image model, the InceptionV3 will be further evaluated in the following combined tasks. It has also been proven that meta-data, with an accuracy of up to 77.8% on the validation set, is an important part of a information disorder detection task. The challenge here is to find meaningful features. In the following multimodal experiments the features of the num_comments and upvote ratio columns will be further evaluated. The next sections will cover multimodal experiments.

### 4.5.2 Dual modal experiments and results

The second part of the experiments will now test all possible combinations of the best single modalities for the two modality approach. Within table 4.10 the assumed parameter settings are presented. Since at this stage feature fusion is required, it has to be noted that at this point the features were fused by only utilizing concatenation as the preferred fusion strategy.

| | Run 1 | Run 2 | Run 3 | Run 4 | Run 5 | Run 6 |
|---|---|---|---|---|---|---|
| Choosen modalities | Title - Visual | Title - Comments | Title - Meta | Comments - Meta | Visual - Meta | Visual - Comments |
| Sequence length | 128 | 128 | 128 | 128 | - | 128 |
| Epochs | 5/10 | 10/10 | 10/10 | 10/10 | 4/10 | 10/10 |
| Batch Size | 228 | 2048 | 4096 | 1024 | 128 | 128 |
| Image Sizes | 768px x 768px | - | - | - | 768px x 768px | 768px x 768px |
| Optimizer - Beta 1 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 |
| Optimizer - Beta 2 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 |
| Optimizer - Epsilon | $10^{-8}$ | $10^{-8}$ | $10^{-8}$ | $10^{-8}$ | $10^{-8}$ | $10^{-8}$ |
| Duration (mean) minutes | 93.80 | 39.2 | 20 | 20 | 86.25 | 90 |
| Accuracy - Validation | 90.8% | 85.9% | 88.1% | 78.2% | 81.1% | 88.0% |
| Accuracy - Test | 91.0% | 85.7% | 88.2% | 78.2% | 81.6% | 88.1% |

*Table 4.10.* Overview about the experiment settings for dual modality.

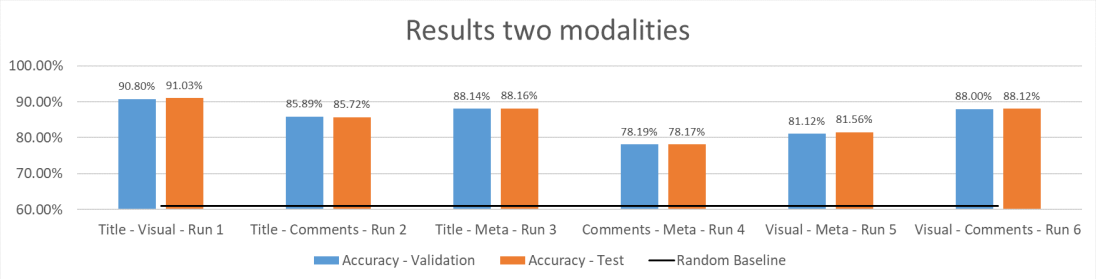


*Figure 4.23.* Results of the dual modality approach.

As it can be seen in table 4.10 and figure 4.23, Run 1 performed the best with a validation accuracy of 90.8% and a test accuracy of 91%. The combination of the reddits title and related images performed very well. As mentioned in the previous section, from now only the Inceptionv3 network is utilized for the visual modality. Also, Run 6 with images and comments performed well and showed clearly the advantages of a multimodal approach. Interestingly, the runs containing meta information (Run 3, 4, and 5) did not improve the accuracy against the mono-modal Run 1 of the title modality. The same picture is shown in Run 5 where the meta-information did not improve the accuracy against the mono-modal visual run, which stays around 88%. Within the next section the three modality approach is introduced.

### 4.5.3 Three modalities

The third part of the experiments will now test all possible combinations of the best single modalities for the three modality approach. Within table 4.11 the assumed network parameters are documented.

| | Title - Visual - Comments | Visual - Comments - Meta | Title - Visual - Meta | Title - Comments - Meta |
|---|---|---|---|---|
| Sequence length | 128 | 128 | 128 | 128 |
| Epochs | 5/10 | 6/10 | 9/10 | 10/10 |
| Batch Size | 96 | 96 | 128 | 96 |
| Image Sizes | 768px x 768px | 768px x 768 px | 768px x 768px | 768px x 768px |
| Optimizer - Beta 1 | 0.9 | 0.9 | 0.9 | 0.9 |
| Optimizer - Beta 2 | 0.999 | 0.999 | 0.999 | 0.999 |
| Optimizer - Epsilon | $10^{-8}$ | $10^{-8}$ | $10^{-8}$ | $10^{-8}$ |
| Duration (mean) minutes | 106.78 | 77.69 | 75.14 | 43.1 |
| Accuracy - Validation | 94.90% | 91.23% | 92.80% | 94.40% |
| Accuracy - Test | 94.99% | 91.30% | 92.80% | 94.46% |

*Table 4.11.* Results and parameters of the three modalities approach.

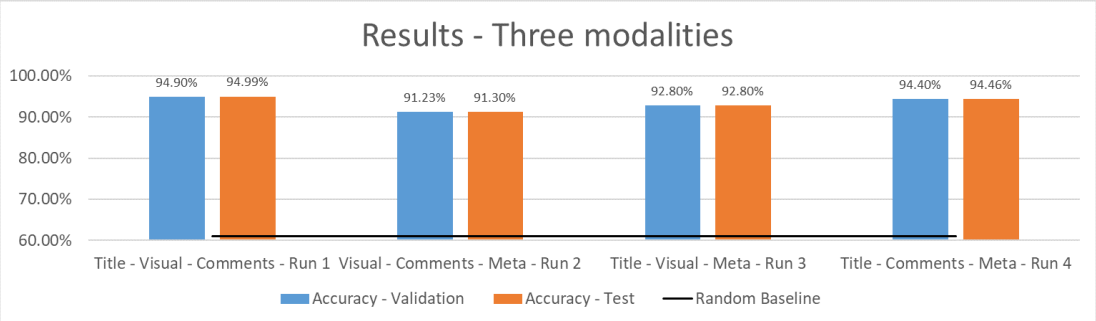


*Figure 4.24.* Results of the three modalities approach.

As it can be seen in figure 4.24 respectively table 4.11 fusing the main modalities title, comments and visual resulted in a very good run validation accuracy of 94,9% and 94,99% on the test set. But with around 106 minutes per epoch training time also the slowest one. Also fusing the modalities title, comments, and meta (Run 4) resulted in a very good validation accuracy of 94,4% and 94,46% on the test set respectively. This run was with 43 minutes per epoch also the fastest one. Only in the case of the last run the maximum count of epochs is reached, this leads to the hypothesis that there could be room for improvement.

Run 1 (Title, Visual, Comments) as a very good run was evaluated by fusing the three modalities by different strategies, namely Concat, Maximum, and Add. The best result, as seen in figure 4.25 was with a minimal better result than Maximum and Add, which was Concatenate with a validation accuracy of 94.90% and 94.99% on the test set. But there is to mention, that the numbers differ only in a very small range of values. There is no longer to speak of a real "improvement". But what should be emphasized is that the best results need the title and comments modalities to reach peak performances. The image and meta-data modality provide only minimal improvement.

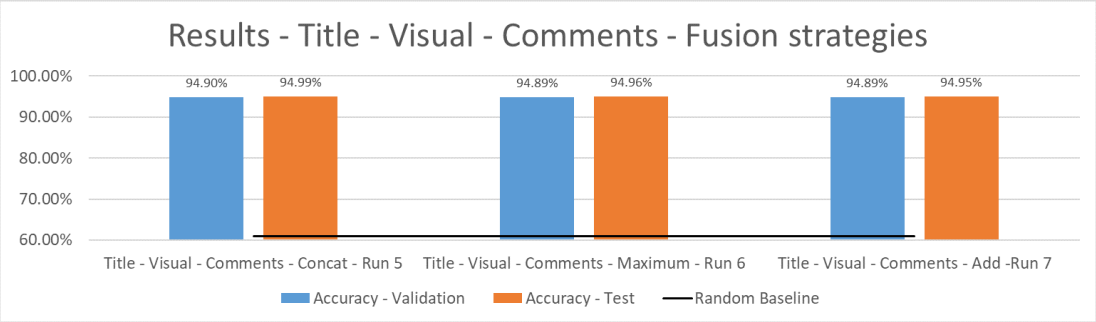


*Figure 4.25.* Results and parameters of the three modalities approach, comparison of different fusion strategies.

The last section of the experiments is going to present the final proposed method, the usage of all four modalities, and its results.

### 4.5.4 Four modalities

The fourth part of the experiments will test the best combination of the single modalities for the four modality approach as seen in the overview figure 3.6. Within table 4.12 the assumed parameter are presented and the results shown. The meta modality was not been fused together with the other modalities. These information parts are concatenated in a later step, the Level 2 Fusion.

| Fusion Method | Concatenation | Maximum | Add |
|---|---|---|---|
| Sequence length | 128 | 128 | 128 |
| Epochs | 6/20 | 8/20 | 15/20 |
| Batch Size | 96 | 96 | 96 |
| Image Sizes | 768px x 768px | 768px x 768px | 768px x 768px |
| Optimizer - Beta 1 | 0.9 | 0.9 | 0.9 |
| Optimizer - Beta 2 | 0.999 | 0.999 | 0.999 |
| Optimizer - Epsilon | $10^{-8}$ | $10^{-8}$ | $10^{-8}$ |
| Duration (mean) minutes | 103.50 | 104.25 | 105.20 |
| Accuracy - Validation | 95.01% | 94.92% | 95.22% |
| Accuracy - Test | 95.20% | 95.10% | 95.54% |

*Table 4.12.* Results and parameters of the four modality approach.

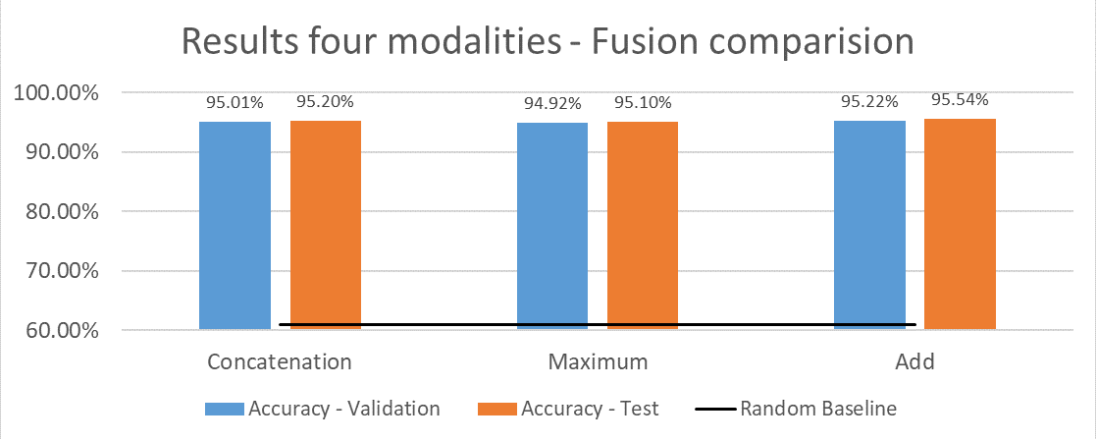


*Figure 4.26.* Results of the model with all four modalities with different fusion strategies.

After the evaluation of the "all modalities" runs an interesting picture emerged. As seen in figure 4.26 and table 4.12 using different fusion strategies slightly different results were achieved. Run 3 was the best run with the Add layer fusion strategy with a validation accuracy of 95.22% respectively 95.54% test accuracy, followed by the Concatenation layer fusion strategy with a validation accuracy of 95.01% respectively 95.20% on the test set. Finally, the Maximum layer fusion strategy was the "worst" one with a validation accuracy of 94.92% respectively 95.10% on the test set. But it has to be noted that the differences between all three fusion strategies are only at a very tiny scale. The results all in all are remarkable and shows the strength of a multimodal approach.

### 4.5.5 Statistical evaluation

With this section some statistical evaluations are going to presented and discussed. The aim is to find out statistical evidence which supports the exoeriments and results of the previous section.

| | Validation Set / Samples | Validation / % | Test Set / Samples | Test / % |
|---|---|---|---|---|
| Number of misclassified samples - Over all models | 157 | 0.27% | 279 | 0.47% |
| Single Modal | | | | |
| Nosw* title was correct but visual was false | 6419 | 10.89% | 24882 | 42.21% |
| Nosw* visual was correct but title was false | 22834 | 38.74% | 3459 | 5.87% |
| Nosw* title was correct but meta was false | 6419 | 10.89% | 11374 | 19.30% |
| Nosw* meta was correct but title was false | 21691 | 36.80% | 5041 | 8.55% |
| Nosw* comments was correct but title was false | 24113 | 40.91% | 5724 | 9.71% |
| Nosw* comments was correct but visual was false | 9158 | 15.54% | 24149 | 40.97% |
| Nosw* comments was correct but meta was false | 7406 | 12.56% | 7600 | 12.89% |
| Nosw* title was correct but comments was false | 3624 | 6.15% | 6651 | 11.28% |
| Nosw* visual was correct but comments was false | 5824 | 9.88% | 3653 | 6.20% |
| Nosw* meta was correct but comments was false | 5671 | 9.62% | 2194 | 3.72% |
| Nosw* visual was correct but meta was false | 9849 | 16.71% | 6522 | 11.06% |
| Nosw* meta was correct but visual was false | 7966 | 13.51% | 21612 | 36.67% |
| Dual Modal | | | | |
| Nosw* single modalities succeeded over best dual modal title visual | 619 | 1.05% | 388 | 0.66% |
| Nosw* single modalities succeeded over best dual modal title meta | 1551 | 2.63% | 89 | 0.15% |
| Three modalities | | | | |
| Nosw* single modalities succeeded over best tripple model | 75 | 0.13% | 115 | 0.20% |
| Nosw* triple model succeeded over all single models | 370 | 0.63% | 90 | 0.15% |
| Four modalities | | | | |
| Nosw* single modalities succeeded over all-four-model | 96 | 0.16% | 52 | 0.09% |
| Nosw* all-four-modal succeeded over single models | 379 | 0.64% | 52 | 0.09% |
| Nosw* all-four-modal incorrect | 2819 | 4.78% | 2739 | 4.65% |
| Nosw* all-four-modal correct | 56125 | 95.22% | 56205 | 95.35% |

*Table 4.13.* Evaluation of the statistical analysis of the model results.

As it can be examined in table 4.13[22] just around 0.27% of the samples of the validation set and 0.47% of the samples of the test set were in total misclassified overall evaluated models. A very interesting picture is revealed if the samples are count which are correctly classified if the title model is considered but falsely detected if only the visual model is taken into account, which number lies around 24882 samples respectively 42% on the test set but only around 11% on the validation set. This leads to the conclusion that the title modality is more meaningful than the image modality. A similar conclusion can be drawn if the comments modality is taken in comparison to the visual modality which is in around 16% of the validation set respectively 42% of the test set misclassified.

The meta modality was a crucial part of this dataset. Looking at the results if the title model predicted the labels correctly, in around 11% of the samples of validation set, respectively around 19% on the test set the meta modality predicted the wrong label. A similar picture can be drawn if the comments modality is considered in opposition to the count of misclassified samples of the meta modality, which is around 13% on the validation and test set.

A very interesting picture can be drawn if the combined model "title - visual" is considered. Only in around 1% of the samples on the validation, respectively 0.66% on the test

[22] *Nosw = Number of samples where ...

set, all single modalities succeeded over the dual-modal approach. This proves that in around 99% of the cases the dual modal approach is better than every single modal approach. A similar evaluation was made if the "title meta" model is considered. There, in 2.63% of the samples on the validation set, respectively 0.15% on the test set the single modalities succeeded over the combined model. A similar picture can be drawn if the best triple models [23] are evaluated.

Considering all four modalities in 99% of all cases the multimodal approach succeeded over all single modalities in combination. So it can be clearly stated, that fusing the knowledge of the single modalities into a multimodal approach is the best way for tackling the problem of information disorder detection.

[23] The model consists out of the title, comments, and visual modality.

## 4.6 Discussion

In this section the results are going to be presented in a more compact layout.

| Type | Val Acc | Test Acc |
|---|---|---|
| Single Modal | | |
| Run Title (Run 3) | **88.10%** | **88.10%** |
| Run Comments (Run 3) | 86.70% | 86.50% |
| Run Visual (Run 5) | 81.04% | 81.51% |
| Run Meta (Run 8) | 77.80% | 77.34% |
| Dual Modal | | |
| Title - Visual (Run 1) | **90.80%** | **91.00%** |
| Title - Comments (Run 2) | 85.90% | 85.70% |
| Title - Meta (Run 3) | 88.10% | 88.20% |
| Comments - Meta (Run 4) | 78.20% | 78.20% |
| Visual - Meta (Run 5) | 81.10% | 81.60% |
| Visual - Comments (Run 6) | 88.00% | 88.10% |
| Triple Modal | | |
| Title - Visual - Comments | **94.90%** | **94.99%** |
| Visual - Comments - Meta | 91.23% | 91.30% |
| Title - Visual - Meta | 92.80% | 92.80% |
| Title - Comments - Meta | 94.40% | 94.46% |
| Four Modal | | |
| Title - Visual - Comments - Meta | **95.22%** | **95.54%** |

*Table 4.14.* Overview about the best models evaluated.

As it can be seen in table 4.14 the results were very promising. All modalities such as title, comments, visual, and meta are clearly over the random baseline of 61%. The most meaningful modality is the title modality which has an accuracy of 88.10% on the validation and test set. The second best modality is the comments modality, followed by the visual and meta modality. Remarkable was, that considering only the best combination of the meta-information, as described in section 4.5.1, the detection accuracy was around 78%.

From the dual-modal approaches, the modalities title and visual was clearly the best combination with an accuracy of around 91% on the validation and test set. The same picture could be taken if the three modalities approach were considered. The best combination was the title, visual, and comments, followed by the title, comments, meta approach with an accuracy of nearly 95% on the validation and test set respectively around 94.5%.

Fusing all four modalities showed that considering the meta modality also improves the accuracy to the best result of all experiments, namely 95.22% on the validation and 95.54% on the test set.

### 4.6.1 Comparison to the state-of-the-art

It is hardly possible to compare the results of this four modality approach with other mono-modal approaches on other datasets. Even if the comparison would be possible on a single modality level, it has also to be considered that the purpose of the method is the same. As there are no other papers using this dataset yet, only a comparison with the authors' (Nakamura et al. 2020) paper can be made.

| Method | Validation (Nakamura et al. 2020) | Test (Nakamura et al. 2020) | Validation thesis | Test thesis | Method |
|---|---|---|---|---|---|
| BERT | 86.54% | 86.44% | **88.10%** | **88.10%** | **BERT** |
| ResNet50 | 80.43% | 80.70% | **81.04%** | **81.51%** | **Inceptionv3** |
| Bert(Title) + ResNet50 | 89.29% | 89.09% | **90.80%** | **91.00%** | **BERT (Title) + Inceptionv3** |
| Add | 85.51% | 85.51% | - | - | - |
| Maximum | 89.29% | 89.09% | - | - | - |
| Concatenate | 85.64% | 85.68% | - | - | - |
| - | - | - | **94.90%** | **94.99%** | **BERT (Title, Comments) + Inceptionv3** |
| - | - | - | 95.01% | 95.20% | Four modalities (Concatenate) |
| - | - | - | **95.22%** | **95.54%** | **Four modalities (Add)** |
| - | - | - | 94.92% | 95.10% | Four modalities (Maximum) |

*Table 4.15.* Comparison with results of (Nakamura et al. 2020).

As it can be seen in table 4.15 the results of the thesis are comparable to the results of Nakamura et al. (2020). The monomodal approaches of the title and image modality are comparable. Due to the fact that the parameter of sequence length of (Nakamura et al. 2020) is not known, this could be the reason that the results of the proposed method of this work are better. The visual modality of both papers are nearly identical. Starting with the three modalities approach the strength of fusion multiple modalities emerges with a leap in accuracy of almost 5%. Utilizing different fusion strategies did not result in the same increase of accuracy as in Nakamura et al. (2020) paper.

So, from the state-of-the-art analysis, it could be clearly stated that multimodal approaches are better than mono-modal approaches. The same conclusion can be confirmed by the experiments in this thesis.

# 5 Conclusion & Future Work

## 5.1 Conclusion

This thesis presented a practical approach to information disorder detection. Firstly by a state-of-the-art literature research, the theoretical background was explored for developing a new method based on it. It was clearly stated that a multimodal approach outperforms mono-modal approaches. The main problem was that no big dataset was available until the Fakeddit dataset was firstly published in 2019. Containing around 640k multimodal samples, this set was a good starting point to perform fake information detection on a large scale level. Secondly, using state-of-the-art text processing methods, such as BERT for the text modalities and well-known image classification networks, such as ResNet50v2, ResNet101v2, and Inceptionv3, allowed to build up a powerful stack of models for the mono-modal and multimodal experiments. After calculating the random baseline, which was around 61%, every single modality was evaluated to find the best performing models. Afterwards the best models (also considering time / accuracy rations) were taken for fusing different modalities, up to the combination of all four modalities, namely title, comments, visual, and meta data. The dataset allowed to evaluate the models up to a 6-label problem. Due to the complexity of the models and topic, only a 2-label setting was evaluated in the scope of this work.

The whole pipeline was implemented by using Functional API from the Deep Learning platform Keras. A pre-analysis made on the dataset in advance showed that, especially for the BERT model, picking the correct sequence length can save a lot of time, by not loosing too much accuracy. Each model was trained and evaluated in a well documented experimental setup. It could be clearly shown which modality and combinations of modalities perform better or worse.

To summon up, the results were very promising and showed the advantages of each modality, but also the powerfulness of fusing different modalities, which resulted in an accuracy of up to 95%. To answer the first research question, which modality is more meaningful, is the answer multi-layered. If only single modalities are considered, the textual modalities are most meaningful. If multimodal models are taken into account, then fusing all of the modalities showed the best results. The combined models showed clearly that choosing the right combinations, had a great impact on the results of information disorder

detection. The best dual modality approach was the title, visual approach, followed by the modalities title, meta-data, followed by comments and visual. The best triple combination of modalities included modalities title, visual, and comments, followed by title, comments, and meta-data. Combining all four modalities performed best, this answers the second research question, to what extent can combined multimodal analysis improve the detection of information disorder. Using state-of-the-art methods, such as a pre-trained BERT-Model and pre-trained InceptionV3 in combination with a suitable meta-data model resulted in the best runs and this answers the third research question, which network architectures from research are best suited for the multimodal analysis of information disorder. It must be noted, that, as described in chapter 4, the meta-data model must be fused at a point, where the value of the features are most meaningful. It could be shown that nearly all available information are extremely useful for being processed by a deep learning neural network.

## 5.2 Limitations and Future Work

Using only multimodal samples from the dataset rather than all reduced the complexity of the problem but also reduced the count of the available samples by a third. Also considering only two labels instead of six[1] made the development of the method easier. So one point for the future work would be on the one hand collecting more samples to tackle the problem of imbalanced classes or develop/find a solution for handling problems with this kind of imbalanced dataset.

A second limitation is that the pre-processing and this method in general is for now only suitable for the use on Reddit itself. Data from other social media platforms such as Facebook or Twitter can not be processed due to the different data formats and available data. Developing a method that can handle data from different platforms or creating a robust method if one of the features is missing is definitely a problem that should be solved by developing appropriate methods.

A third limitation is the general problem of information disorder detection. It is questionable that a sample always belongs to only one class. The proposed method can only handle one label per sample and not multiple labels per sample. Additionally overcome the concept of just fake or non-fake to more fine-grained analysis of different sub-types of fake information. Furthermore a tool could be developed to highlight posts, or whole news articles, or parts of the text, or image that could be identified as fake or manipulated content.

Another possibility for future work is using explainability methods to explain which samples are wrongly or truly classified and why. So the next generation networks can be better designed for future tasks.

Finally, a more in-depth analysis could be performed by evaluating different models for the same modality and if different methods on the same modality can improve the detection result. This can be applied to all modalities.

To sum it up, information disorder detection is and will remain a challenge for our society. The effects of information disorder, i.e. the problems of social, political and technical nature will accompany mankind on many levels in its daily life and technical systems. Methods, as for example presented in this work, can help to tackle the problem, but will never replace common sense and critical questioning of information.

[1] The development of a method for the six label problem is much more difficult because of the fact that the classes are heavily imbalanced.

# Bibliography


Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, *31*(2), 211–236. doi:10.1257/jep.31.2.211

Boididou, C., Papadopoulos, S., Zampoglou, M., Apostolidis, L., Papadopoulou, O., & Kompatsiaris, Y. (2018). Detection and visualization of misleading content on Twitter. *International Journal of Multimedia Information Retrieval*, *7*(1), 71–86. doi:10.1007/s13735-017-0143-x

Bridle, J. S. (1990a). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing* (pp. 227–236). Springer.

Bridle, J. S. (1990b). Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In *Advances in neural information processing systems* (pp. 211–217).

Buntain, C., & Golbeck, J. (2017). Automatically Identifying Fake News in Popular Twitter Threads. *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, 208–215. arXiv: 1705.01613. doi:10.1109/SmartCloud.2017.40

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Cozzolino, D., Poggi, G., & Verdoliva, L. (2015). Splicebuster: A new blind image splicing detector. In *2015 IEEE International Workshop on Information Forensics and Security (WIFS)* (pp. 1–6). IEEE.

Cui, L., Wang, S., & Lee, D. (2019). SAME: Sentiment-Aware Multi-Modal Embedding for Detecting Fake News, 8.

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). Botornot: A system to evaluate social bots. In *Proceedings of the 25th international conference companion on world wide web* (pp. 273–274).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). Ieee.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, M., Yao, L., Wang, X., Benatallah, B., Sheng, Q. Z., & Huang, H. (2018). Dual: A deep unified attention model with latent relation representations for fake news detection. In *International Conference on Web Information Systems Engineering* (pp. 199–209). Springer.

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, *12*(7).

Fallis, D. (2014). A functional analysis of disinformation. *iConference 2014 Proceedings*. Publisher: iSchools.

Farid, H. (2009). Exposing digital forgeries from JPEG ghosts. *IEEE transactions on information forensics and security*, *4*(1), 154–160. Publisher: IEEE.

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104. Publisher: ACM New York, NY, USA.

Ferrara, P., Bianchi, T., De Rosa, A., & Piva, A. (2012). Image forgery localization via fine-grained analysis of CFA artifacts. *IEEE Transactions on Information Forensics and Security*, *7*(5), 1566–1577. Publisher: IEEE.

Forelle, M., Howard, P., Monroy-Hernández, A., & Savage, S. (2015). Political bots and the manipulation of public opinion in Venezuela. *arXiv preprint arXiv:1507.07109*.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249–256).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Gruppi, M., Horne, B. D., & Adalı, S. (2020). NELA-GT-2019: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles. *arXiv preprint arXiv:2003.08444*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016a). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). doi:10.1109/CVPR.2016.90

He, K., Zhang, X., Ren, S., & Sun, J. (2016b). Identity mappings in deep residual networks. In *European conference on computer vision* (pp. 630–645). Springer.

Heller, S., Rossetto, L., & Schuldt, H. (2018). The PS-Battles Dataset - an Image Collection for Image Manipulation Detection.

Hernon, P. (1995). Disinformation and misinformation through the internet: Findings of an exploratory study. *Government information quarterly*, *12*(2), 133–139. Publisher: Elsevier.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780. Publisher: MIT Press.

Horne, B. D., Khedr, S., & Adali, S. (2018). Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *Twelfth International AAAI Conference on Web and Social Media*.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Howard, P. N., & Kollanyi, B. (2016). *Bots, #Strongerin, and #Brexit: Computational Propaganda During the UK-EU Referendum* (SSRN Scholarly Paper No. ID 2798311). Social Science Research Network. doi:10.2139/ssrn.2798311

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Jin, Z., Cao, J., Guo, H., Zhang, Y., & Luo, J. (2017). Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. *Fast Forward*, 9.

Kim, J., Tabibian, B., Oh, A., Schölkopf, B., & Gomez-Rodriguez, M. (2018). Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (pp. 324–332).

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

Kingma, D. P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*. arXiv: 1412.6980. Retrieved July 1, 2020, from http://arxiv.org/abs/1412.6980

Kochkina, E., Liakata, M., & Augenstein, I. (2017). Turing at semeval-2017 task 8: Sequential approach to rumour stance classification with branch-lstm. *arXiv preprint arXiv:1704.07221*.

Kumar, S., & Shah, N. (2018). False Information on Web and Social Media: A Survey. *arXiv:1804.08559 [cs]*. arXiv: 1804.08559. Retrieved February 20, 2020, from http://arxiv.org/abs/1804.08559

Lago, F., Phan, Q.-T., & Boato, G. (2019). Visual and Textual Analysis for Image Trustworthiness Assessment within Online News. Research Article. doi:https://doi.org/10.1155/2019/9236910

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

Liu, X., Nourbakhsh, A., Li, Q., Fang, R., & Shah, S. (2015). Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (pp. 1867–1870).

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., & Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. In *Ijcai* (pp. 3818–3824).

Ma, J., Gao, W., & Wong, K.-F. (2017). Detect rumors in microblog posts using propagation structure via kernel learning, Association for Computational Linguistics.

Ma, J., Gao, W., & Wong, K.-F. (2018). Rumor detection on twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1980–1989).

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

Mitra, T., & Gilbert, E. (2015). CREDBANK: A Large-Scale Social Media Corpus With Associated Credibility Annotations. In *ICWSM* (pp. 258–267).

Mohtarami, M., Baly, R., Glass, J., Nakov, P., Màrquez, L., & Moschitti, A. (2018). Automatic stance detection using end-to-end memory networks. *arXiv preprint arXiv:1804.07581*.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *ICML*.

Nakamura, K., Levy, S., & Wang, W. Y. (2019). R/Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. *arXiv:1911.03854 [cs]*. arXiv: 1911.03854 version: 1. Retrieved December 18, 2019, from http://arxiv.org/abs/1911.03854

Nakamura, K., Levy, S., & Wang, W. Y. (2020). R/Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. *arXiv:1911.03854 [cs]*. arXiv: 1911.03854. Retrieved April 11, 2020, from http://arxiv.org/abs/1911.03854

Nørregaard, J., Horne, B. D., & Adalı, S. (2019). NELA-GT-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 13, pp. 630–638).

Pathak, A., & Srihari, R. (2019). BREAKING! Presenting Fake News Corpus for Automated Fact Checking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop* (pp. 357–362). doi:10.18653/v1/P19-2050

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Peters, M. E., Neumann, M., Zettlemoyer, L., & Yih, W.-t. (2018). Dissecting contextual word embeddings: Architecture and representation. *arXiv preprint arXiv:1808.08949*.

Pomerleau, D., & Rao, D. (2017). *Fake news challenge*.

Ruchansky, N., Seo, S., & Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 797–806). ACM.

Salem, F. K. A., Feel, R. A., Elbassuoni, S., Jaber, M., & Farah, M. (2019). FA-KES: A Fake News Dataset around the Syrian War. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*, 573–582. Retrieved July 10, 2020, from https://aaai.org/ojs/index.php/ICWSM/article/view/3254

Santia, G. C., & Williams, J. R. (2018). Buzzface: A news veracity dataset with facebook user commentary and egos. In *Twelfth International AAAI Conference on Web and Social Media*.

Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., & Menczer, F. (2017). The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592*, *96*, 104. Publisher: ArXiv e-prints.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, *9*(1), 4787. arXiv: 1707.07592. doi:10.1038/s41467-018-06930-7

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2019). FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media. *arXiv:1809.01286 [cs]*. arXiv: 1809.01286. Retrieved March 5, 2020, from http://arxiv.org/abs/1809.01286

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*. arXiv: 1409.1556. Retrieved March 31, 2018, from http://arxiv.org/abs/1409.1556

Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., & Satoh, S. (2019). SpotFake: A Multi-modal Framework for Fake News Detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)* (pp. 39–47). doi:10.1109/BigMM.2019.00-44

Skansi, S. (2018). *Introduction to Deep Learning: From logical calculus to artificial intelligence*. Springer.

Smith, L. N. (2017). Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 464–472). IEEE.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958. Publisher: JMLR. org.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). Rethinking the Inception Architecture for Computer Vision. *arXiv:1512.00567 [cs]*. arXiv: 1512.00567. Retrieved July 17, 2020, from http://arxiv.org/abs/1512.00567

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). Some Like it Hoax: Automated Fake News Detection in Social Networks. *arXiv:1704.07506 [cs]*. arXiv: 1704.07506. Retrieved July 10, 2020, from http://arxiv.org/abs/1704.07506

Tan, M., & Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism quarterly*, *30*(4), 415–433. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 809–819). doi:10.18653/v1/N18-1074

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. Publisher: American Association for the Advancement of Science.

Wang, W. Y. (2017). " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., ... Gao, J. (2018). Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 849–857). ACM.

Wardle, C. (2017). Fake news. It's complicated. *First Draft*, *16*.

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe report*, *27*.

Zach, F., Riess, C., & Angelopoulou, E. (2012). Automated image forgery detection through classification of JPEG ghosts. In *Joint DAGM (German Association for Pattern Recognition) and OAGM Symposium* (pp. 185–194). Springer.

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, *57*(2), 102025. doi:10.1016/j.ipm.2019.03.004

Zubiaga, A., Liakata, M., & Procter, R. (2017). Exploiting context for rumour detection in social media. In *International Conference on Social Informatics* (pp. 109–123). Springer.

Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., & Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, *11*(3), e0150989. Publisher: Public Library of Science San Francisco, CA USA.

# List of Figures

# List of Tables