Project abstracts

Classifying subreddits for Reddit posts using Naïve Bayes

This study tried to classify the appropriate subreddit given a certain Reddit post by utilizing a Naive Bayes classifier. The data consisted of approximately 1 million posts with 1013 different subreddits. Initial research found that work in the area was rather limited, and an exploratory approach was adopted, testing different methods and evaluating the results. The main research question was whether certain features could be incorporated into a Naive Bayes classifier, drawing on the information provided in the posts and improving the classifier's performance. The data contained the title of the post as well as the body, i.e. the actual post. Using spaCy, part-of-speech tags were generated for the dataset to try and create morphological features that might be predictive for certain subreddits. The model outperformed another classifier found on Kaggle, but did not so on account of the features engineered, but rather by preprocessing the data differently. The model showed a 89.6% precision at finding the correct post in its top five predictions, a 1% increase over the Kaggle counterpart.

Myers-Briggs personality type prediction with text classifier

Myers Briggs Type Indicator (MBTI) is an assessment that divides personality types over four major binary axis defined as (Introversion, Extroversion), (Sensing, Intuition), (Feeling, Thinking) and (Perceiving, Judging). The combination of these indicators make up 16 unique types that are widely used to sort people with respect to their dominant psychological functions. The types have proven to be highly correlated with how one makes decisions or takes action in personal and professional contexts. This report aims to explore the underlying patterns and sentiment within a collection of user posts, labeled with their respective personality type, and build a classifier that can be applied to determine personality type of a person based on text he/she has produced. This report attempts to study a number of re-sampling techniques as a pre-step, evaluate classifiers with different parametrizations and implement an LSTM model to build a classifier with high-accuracy. Final model is compared to results of a similar paper by Stanford University and tested on Donald Trump's twitter data.

Predicting personality types of 'Song of Ice and Fire'-characters

Not all prospective models have access to training data that is suitable for the task at hand, and at times requires the training to be conducted on data that is not obviously representative of the data that the model will be used for. As such, this report investigates the feasibility of using social media posts as a basis for training a classifier that attempts to predict personality types of characters from a fictional books series. A logistic regression model will be used for classification along with count vectors as features. Both a count vectorizer and tfidf vectorizer is tested, and the logistic regression model's parameters is fined tuned with grid search. The result showed poor performance for predicting the character's personalities, most likely due to the complexity of the task and unbalanced training data. As such, a second test was performed where the complexity was the reduced and the data better balanced. The results showed an increase in performance and indicated that it may be feasible to use social media posts for training in order to predict fictional character's personality.

Model Comparison for Review Classification

There are a wide variety of classification models available but the hardest part is trying to pick which one to use for the task at hand. Depending on the type of data, hardware resources, amount of data and other factors this can vary a lot. This project is trying to find answer to how available resources and data volume influence this decision between a number of Machine Learning models and a RoBERTa Transformer. While clearly outperforming the other models on little data the Transformer couldn't be trained on higher amounts of data. Because of this it was surpassed by not just the moderately time consuming Support Vector Classifier but also the very efficient Multinomial Naive Bayes.

Keyword Extraction from Swedish Job Ads

Employers are increasingly using information technology, such as Applicant Tracking Systems (ATS) to aid in recruiting (Laumer, Maier and Eckhardt 2015). Meanwhile, patents are being filed for automated filtering of job applications, based on keywords (Shah 2020, Dey 2012). Increasingly job applicants will need to not just list their relevant experiences, they need to do so using the correct words (Zahn 2018, Novak 2017). This report tries to develop a suitable keyword extractor to aid job applicants write their applications. For the keyword extraction two main methods are tested: term frequency - inverse document frequency (tf-idf) and a manually defined set of rules based on Part-of-Speech (POS) tagging and dependency parsing of sentences from job applications. The tf-idf approach resulted in the best recall (recall = 0.7341) of the two methods but had a low precision (precision = 0.2768). The rules-based approach was able to perform best overall (recall = 0.5209, precision = 0.5080, f1 = 0.5144).

Genre Classification Based on Song Lyrics

In this paper text classification of song lyrics has been studied with the goal of classifying lyrics into music genre classes. The genre classes were defined by reviews from the online music publication Pitchfork. The studied classifiers were classifiers based on count and tf-idf vectorizers and Multinomial Naive Bayes, Multilayer Perceptron Classifier and Support Vector Machine Classifier predictors. All of the studied classifier pipelines outperformed baseline results, which were two baseline classifiers using the 'stratified' and 'most frequent' sampling strategies. The Support Vector Machine Classifier was analysed closer and an exhaustive grid search cross-validation was conducted on this classifier to tune its hyperparameters, which increased its performance. From the results of the tuned classifier it was observed that the F1 score for the rap genre was significantly higher than the second highest F1. Further analysis revealed that the rap genre had the largest amount of total words and unique words per lyric on average among all studied genres. Furthermore, investigating the most significant features of a linear Support Vector Machine Classifier revealed that the most significant features when classifying rap were very unique, genre specific features. In conclusion rap is arguably the most lyrical among the studied genres.

Utilizing NLP transfer learning with ULMFiT in Chemoinformatic

This project explores the capabilities of NLP transfer learning on a text representation other than a natural language. In chemoinformatics, the molecules can be represented as strings of characters (SMILES) and they can be treated as a 'chemistry language'. The Universal Language Model Fine-tuning for Text (ULMFiT) is used, pre-trained on an unlabeled big dataset of one million molecules to learn the 'chemistry language'. Then this model is fine-tuned to predict the lipophilicity property of a molecule which is a continuous value. The pre-training part of the model takes around 6 hours and predicts correctly 82% of the time the next token given the previous 50 tokens. After the model is fine-tuned to predict the lipophilicity of a molecule and it achieved root mean squared error (RMSE) 0.79. While a base line model like random forest which uses features by feature engineering from the molecules, achieves RMSE around 0.98. The results shown that using NLP transfer learning on chemoinformatics achieves strong performances that can outperform the models on structured data by feature engineering.

Sentiment Analysis and Classification of Twitter Data about General Elections Held in India

Nowadays politics is a key issue. It affects the general feeling of whole countries and people express their opinion through social networking platforms such as Facebook, Twitter or Instagram. Therefore, this project focuses on the implementation of different models to classify Tweets depending on the feeling contained in their words. To this purpose, it has been used a database, obtained from Kaggle, that contains 163K tweets written during the general elections in India. Different types of processing, feature extraction and 5 models have been evaluated. These models include 2 based on neural networks in addition to the other 3 that have been studied during this course. The data have been classified in 3 classes: negative, neutral and positive. The model that presents the best results is the LSTM, with which an accuracy of 83.69% has been achieved but there are no notable differences with others. Despite the fact that it shows good results, the network is not able to generalize with others tweets that don't belong to this theme.

Automatic Text Summarization - Evaluation of Seq2seq LSTM & Seq2seq Attention

Abstract Aim of the automatic text summarization is to extract or abstract the information which expresses the main idea of the document. Seq2seq is a framework which can be used for text summarization. Currently we have abundant of models available online but we don't have good comparison available for the better understanding. In this report we have targeted to compare two models based on seq2seq LSTM and seq2seq attention and evaluate how both models are behaving under different experiments. At the end of the report we will talk about the results of the each experiment as well as in the discussion we will state about some possibility how this work can be extended in future.

Classification of song genre by lyrics

This paper will focus on the hypothesis that models with "memory", i.e. take into account the word order, are preferred when predicting the song genre where you only use the lyrics. Hence, the ordering of the words provides additional information to the predictions. This was made possible by using the dataset "Spotify music genre list and 80k songs/tracks" Molina (2018) I was able to receive music genre from each song, but unfortunately not the lyrics. Using the combinations between Spotify API and Genius lyrics API there was possible to retrieve the correct lyric to the matching song. The paper will compare three models, Gated Recurrent Unit (GRU), Long Short-Term

Memory (LSTM) as models with memory and the Multinomial Naive Bayes classifier (MNB) as model without memory. The results shown that both the GRU and LSTM both accuracy is close to 0.36 and the MNB had a accuracy close to 0.37. Given the data material and the selected models for the study, we can assume that when it comes to predicting which music genre, the ordering of the words may not have a major impact, but which words used have the greatest impact on the predictions.

Adversarial attacks on e-mail spam classifiers

Machine learning has become a reality in many day-to-day applications, one of them is spam classifying. Spam e-mails constitute a 30% of the total traffic on the internet and signify a big problem to companies; therefore, they must be correctly filtered. The question that arises is, are these algorithms infallible? Adversarial attacks on machine learning algorithms have proved that the answer is no. By slightly modifying the inputs of the algorithms, they achieve to fool the classifiers. In this report, different algorithms were tested for the spam e-mail classifying task. Then, three attacks have been implemented and used against the classifiers to compare their behaviours. The evaluation of the methods has been measured by means of accuracy and F1-score.

Natural Language Inference on a small Multilingual Dataset

The goal of this project was to explore different methods of natural language processing and investigate how they compare on a small and challenging dataset in the task of natural language inference (NLI). We investigate different methods and see how they perform in this scenario. We explore simpler methods of neural networks trained with manual features as well as word embeddings, recurrent neural networks and finally transformer architectures. We also explore the topic of pretraining in this context since the used dataset is relatively small. Another challenging aspect of the dataset is that it does not only contain English samples but samples from 15 different languages. We tackle this problem both by using translation as well as multi-lingual models and compare both approaches.

Mood classification of songs through lyrics analysis - a text mining approach

This paper presents a series of experiments with machine learning and mood classification of English song lyrics. It covers the process of extending an existing lyrics dataset with mood gold labels. It then explores the impact of different methods of feature vectorization and classification algorithms on the precision and recall of the classification task at hand. The experimental results show a better performance for the Multi-Layer Perceptron Classifier with a precision of 56%. Potential caveats and areas of improvement are also discussed. Keywords: lyrics, music mood classification, machine learning, web scrapping.

Using LSTMs to Generate Answers to Stack Overflow Questions

Natural Language Generation (NLG) is a known field within Artificial Intelligence and Natural Language Processing (NLP). As the name of NLG suggests, the field covers all types of natural language generation, meaning any model that tries to produce human-like text, that is understandable to humans. This project investigates if it is possible to produce a human-like answer to a Stack Overflow question using an LSTM model. To evaluate the generated answer the ROUGE-1 and ROUGE-2 precision was used, which calculates the number of overlapping unigrams and bigrams between a generated answer and the gold standard answer. The best model implemented was able to generate answers that resulted in 21.99% precision for the ROUGE-1 metric, but only 0.5% precision for the ROUGE-2 metric.

Sentiment analysis with EDA

More often than not training data is insufficient for training a model, or there is an imbalance in the data. This text mining project aims to explore and evaluate the performance of easy data augmentation (EDA) on the IMDB movie review set. The data augmentation is applied using a method called Synonym Replacement in three different ways. A baseline accuracy will be measured using Logistic regression and Multinomial Naive Bayes. The augmented data is later evaluated using Bidirectional Encoder Representations from Transformers (BERT), to see if augmentation of data increases the accuracy.

Semantic similarity on Quora question pair dataset

The project explores the task of semantic analysis by looking at Quora Question pair DataSet. We conducted intense cleaning of data and various idea for feature extraction and used different type of deep learning models with different word embedding. Our final finding was LSTM with attention layer using Glove word embedding is a best model in term of accuracy or loss and computational power. We achieved maximum of 84% accuracy on simple Bi-LSTM model.

Identifying actionable app reviews

In this project, three methods for classifying reviews as relevant or not based on the review content was devised and tested. One method used TF-IDF vectorization and Nearest Neighbor for classification. The second and third method both utilized different configurations of a Multi-layer Perceptron classifier. Each method was evaluated using precision, recall and F1 score. The results were then compared to those of a ”random guess” baseline. Compared to the baseline, the second and third method both improved the F1 score, but the first method did not. The data set used for analysis contained information about ”apps” from the Shopify apps marketplace and was provided by the data service Kaggle. A Gold standard with 499 entries was created from randomly sampling the original data set. This Gold standard was used for both training and testing using a 80/20 split.

Guess the gender - A closer look on book genre prediction based on reader descriptions

The goal of this project was to classify which book that belonged to which genres, based on book descriptions from readers on the webpage Goodreads. To solve this problem three different approaches were compared and investigated: Multinominal Naïve Bayes, LSTM and BERT. Evaluation of the performance of the machine learning models is based on the metrics precision, recall, accuracy and F1-score. Out of the models used in this study, BERT achieved the highest performance.

Produce Code Descriptions from Source Code

The ability to translate between source code and natural language has tremendous applications and assisted coding could help programmers and non-programmers alike to produce meaningful programs. Even though this project is to small to do exactly that, this project have focused on the ability to produce a description of a code given nothing but the source code free from docstrings and comments. The model that have been used is the transformer architecture proposed in the paper *Attention is all you need*. In order to evaluate the results, BLEU-score and n-gram precisons were used. Unfortunately, the obtained results were a mere BLEU-score of 0.02, an 1-gram precision of 22.4 % and a 4-gram precision of 0.22 %. Despite to scores, the actual translation shows that it is possible to obtain meaningful description of code given only its source.

Sentiment on movie reviews using LSTM and Bidirectional-LSTM

Using movie reviews to choose which movie to watch is useful but can be time consuming, would it not be easier to use a machine learning model to do this for us? This is what is evaluated in this project, if using a LSTM is a good model to predict the sentiment of a movie based on its reviews. Additionally it is tested if using a bi-directional LSTM model improves the model and also which works best, pre-trained or task specific embedding matrices. The dataset used is 50000 imdb movie reviews from the site kaggle. The evaluation of these questions are done by measuring their accuracy and comparing them to each other and to a baseline model. The conclusion of this project is that using a bi-directional LSTM model or using a pre-trained embedding with task specific training producers the best model for this task.

Evaluating song emotions impact on song popularity

There are many factors to what makes a song popular. This project aims to find out whether or not the mood of the song contributes to the song's popularity. The mood-features will be predicted from the song lyrics using a trained mood-classifier that trained on a different dataset. The methodology is to compare two different song popularity classifiers. One that only trains and predicts on numerical features such as tempo, duration and energy, while the other model trains and predicts on the mood-features, in combination with the numerical features. This process will be done for a SVM and a logistic regression model. The results were that the mood features did not impact the song popularity and that the accuracy of the two models were slightly higher than random. Finally, the mood classifier had an accuracy of 73%.

A Method to Extract Support Phrases for Sentiment Labels

Maintaining the right impression in the minds of customers is of ut-most importance to any company to prevent churn and gain new customers or following, and their social media presence such as Facebook,Twitter and Instagram are a big chunk of that. Sentiment analysis of posts on social media is a good way to gauge a product's popularity. Correctly identifying the phrases/keywords that showcase a particular sentiment in a sentence is a key part of these large-scale sentiment analysis projects.This study analyses tweets to extract the phrases resulting in the given sentiments of those tweets using BERT transformer model and evaluates the performance of model using the Jaccard similarity metric.

Sentiment analysis on movie reviews using modern NLP models

Sentiment analysis is a common area in NLP, and have many useful applications such as customer review analysis. This project tests two modern transformer-based models for identifying whether the sentiment of a movie review is positive or negative and compares the models to a baseline model made of a tf-idf vectorizer and a logistic regression classifier. The results show that the modern networks made from a transformer network and bidirectional LSTM had lower test accuracy than the baseline model. This means that old methods like tf-idf and logistic regression might still be relevant for sentiment analysis and binary text classification.

Stock Price Prediction through Sentiment Analysis of Stock News

Financial news may have a significant impact on the regarded company's stock price. With articles being published at rapid rates all over the world, it's hard for a human to keep up to date with financial news. This research investigates how different machine learning models perform at classifying news articles into negative, neutral, or positive classes, based on how the article affects the regarded companies future stock price developments. The models that are compared are Multinomial Naive Bayes, Bidirectional long short-term memory networks (BiLSTM), and a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model. Stock prices and financial news articles were collected for the years 2019 and 2020 from a total of 2000 companies. The results show that news articles have the most measurable impact 30 minutes after publication, and the fine-tuned BERT model performed best at classifying the news articles. The model achieved an accuracy of 44.4%, f1-score of 44.0%, and an average rate of return of 0.489%. This research contributes to the research area by investigating how news affects stock prices in a short time frame from publication, based on a new dataset, and how the state of art NLP model BERT performs at the task.

Can /r/WallStreetBets predict the stock market?

/r/WallStreetBets (WSB) is a stock-trading community on the social media platform Reddit. WSB is known for their aggressive trading strategies with heavily leverage options, often trading the SPDR S&P 500 Trust ETF (SPY) index. This study takes a sentiment analysis approach to examine if there is any correlation between what is being said in the comments on the daily discussion thread ”what are your moves tomorrow”, and how the SPY index performs. Three different approaches are implemented and evaluated: multinomialNB, VADER, and LSTM networks. The results indicate that there might be some connection between the sentiment of WSB and the stock market, with the lexicon-based VADER achieving the best results, with 16 % higher sentiment on positive days than negative days. The The most important conclusion from the work is that the dataset used was very noisy, which resulted in the machine learning-based approaches having a hard time learning any usable patterns.

Chatbot for a service company

Since the development of social media networks, more and more businesses have been using chat services to attract more customers by providing real-time information for their potential or existing customers. Medium or large scale businesses have a dedicated customer service person to cater to their online customers' needs. But for small-scale businesses are unable to compete with their rival companies since they have a limited amount of employees and unable to pay for an extra person for online customer handling. And also there are some special service-related businesses in which the owner needs to directly interact with customers. So they can't use an employee to handle their customer because of the business model or they can't expose customer information with another party. Over the last few years, we could see that Text Mining and Natural language processing has been spreading to varies kind of industries. So we could use the Text mining knowledge to address the above situations and create an automated chatbot. We could extract existing customer's text messages from a specific industry and create a simple chatbot that could communicate with customers and solve their problems in real-time.

Extracting Owner Information from Press Releases

This text mining project has made an attempt at extracting insider transactional data from Norwegian press releases. Through the use of spaCy's rule-based matcher, the date at which this insider transaction occurred, the name of the organization or individual who made the transaction, the number of shares that were traded, and the price at which the stocks were either bought or sold at, were extracted from the financial press releases. The project shows that a rule-based matcher using NLP can be used as a baseline for more complex models, or help anyone manually performing the information extraction, and a precision and recall of 0.63 and 0.61, respectively, was reached. Furthermore, the results of this project highlight the importance of knowledge bases and the domain specific knowledge required in order to reach sufficient results, through the use of well-defined rules.

Multi class classification using a recurrent neural network

Hotel reviews are useful when trying to figure out which hotels to stay at. One website that hosts such reviews is TripAdvisor. A review on TripAdvisor consists of a explanatory text, a title and a rating. This report aims to predict the rating for each review based on the text in the review. This type of task is called *multi class classification* and this report uses a recurrent neural networks with long short term memory to predict the ratings for each review. The report aims to explore how well the stated model works to predict the ratings and to see how the number of epochs used during training affects the models final result. The result is a model that is able to predict a correct rating of a review with a accuracy of 58.5% and it is found that a model that is trained over too many epochs leads to over-fitting.

Evaluation of sentiment analysis approaches to the Story Cloze Test

Understanding a simple story is an easy task for human beings but complex for a computer. What determines what is an appropriate ending or next sentence to a given story? In this experiment we use the Story Cloze Test developed by Mohammad et al which consists of a given context of four sentences and the task of determining which out of two given fifth sentences is the correct one. We follow the hypothesis that this can be determined by using sentiment analysis on the given context and the two endings, choosing the ending which best matches the context with regards to sentiment. For this classification two different approaches were used, Naive Bayes classifiers trained on IMDB reviews and one trained on food reviews as well the approach of using the lexical sentiment analysis libraries Afinn and Vader. The results showed that the lexical sentiment analyzers performed better than the Naive Bayes classifiers but that almost all performed better than the sentiment analysis baseline presented in the original paper of 49,2% accuracy. The best result was achieved using a non-binary scoring Vader analyze which resulted in 60% accuracy.

Analyzing Premier League Supporters Opinion on VAR

This work conducts a sentiment analysis on Premier League supporters opinion on VAR. Using data from twitter in the form of tweets, it is investigated if the overall opinion from the supporters is negative or positive. After the tweets have been cleaned and made available for analysis, a model is trained on tweets that have their sentiment manually annotated and are then used to classify the tweets from the Premier League supporters to find the sentiment of each tweet. It is also investigated if the league position of the team and the number of VAR decisions for each team has any impact. The results show that the overall opinion of the supporters regarding VAR is negative and that the combination of a lower league table position and VAR decisions given against a team have an impact on the supporters reaction, although it is not clear which factor has the largest impact. It is concluded that more data from a full season is needed in order to confirm the results.

Genre Classification of Blurbs Using Contextualized Language Models

Pre-trained contextualized language models like BERT have in the last few years caused a paradigm shift in the field of natural language processing (NLP), producing state-of-the-art results on numerous downstream NLP tasks. However, BERT has not yet been extensively adopted for all such tasks, including hierarchical multi-label text classification (HMLTC). HMLTC is the problem of assigning one or more class labels to a text, with the classes being stored in a hierarchical structure. This project examines how BERT can be fine-tuned for HMLTC using the Blurb Genre Collection dataset. Two model architectures based on BERT-BASE are trained and evaluated: the more straightforward BB-BERT and the more experimental HiERBERT, the latter combining a global and local classifier approach. After hyper parameter tuning, these models jointly outperform existing baselines built on CNNs, RNNs and more traditional approaches. BB-BERT is superior across three out of four metrics, indicating that BERT provides a good basis for HMLTC with the capacity to learn class hierarchies. The results for HiERBERT are more inconclusive but nonetheless promising. The project also displays how BERT is fairly sensitive to hyperparameter settings during fine-tuning.

Generating text using LSTM GAN in PyTorch

The aim of the project is to generate text using generative adversarial networks (GANs), where both components are recurrent/LSTM neural networks. The training data set is a collection of internet comments from YouTube, Twitter and Kaggle also containing a (binary) negative sentiment indicator. The model output is evaluated using the result of Markov chain trained on the same data. Absolute quality of the text is evaluated manually against a test set. Sentiment input integration into LSTM GAN is discussed.

Fake Yelp Review Detection using GPT2 and LSTMs

Fake reviews are a prominent issue that affect many businesses. Many customers rely on online reviews to help guide them towards the best restaurant or buy a good product. If a business owner could simply generate or buy fake reviews for their business, portraying an artificial quality, this can lead to manipulation of customers.

In this paper a GPT2 model is fine tuned to be able to generate fake Yelp reviews for restaurants. The fake reviews are used to train different classifiers that decide if a review is fake or real. A LSTM classifier could only achieve accuracy of around 61% on detecting fake reviews. These results show the ability and ease at which fake reviews can be used to sway potential customers into buying someones product or service.

Fine-tuning a Swedish BERT with Federated Learning

Due to regulatory and private reasons, sharing data between entities can often be hindered. Because of this, existing real-world data is not fully exploited by machine learning (ML). An approach to solve this is to train ML models in a decentralized manner using federated learning (FL). In this project, Swedish BERT models are fine-tuned with FL for the task of multi-class document classification. The data set used are speeches retrieved from the Swedish Parliament. Furthermore, the models are compared to a centralized trained model with respect to precision, accuracy, recall, and macro average F1-score. When the training is split into three locations, the results show that models trained with FL give comparable results to a centralized trained model. Using 4 or 5 locations, a slight performance drop was observed. The decrease in F1-score performance ranged from 0%-6%. It is concluded that FL could be a viable choice for training models with decentralized sensitive data.

Evaluation of different classification algorithms for COVID-19 Pandemic Tweets

COVID-19 pandemic has drastically changed the daily lives and routine of people around the world. These changes have affected both the mental health and the emotions of the people who often post about how they feel in social media, such as Twitter. Their posts may express anger, sadness, frustration, optimism, or sometimes nothing at all. Thus, it is of interest to analyze and classify these posts based on the type of sentiment. Several classification methods were used, which among others are Logistic Regression, Multinomial Naive Bayes, Linear Support Vector, Decision Tree, Random Forest, Extreme Gradient Boosting, Stochastic Gradient Descent and Recurrent Neural Network. The performance of the models was evaluated on their accuracy of predicting correctly the type of sentiment on unseen data. Most of the classifier models performed the same without much difference between them, however Decision Tree Classifier and Multinomial Naive Bayes performed the worst. The classifier that outperformed every other one is Recurrent Neural Network classifier, which achieved a really high accuracy compared to the rest of them.

Analysis and Evaluation of Multi label Movie Review Classification using Pre Trained BERT Model and Baseline Classification Techniques

State of the Language Models like BERT, GPT 2, XLNET, T5 are often fine tuned for basic downstreaming classification tasks in NLP. In this report, I experiment the accuracy and efficiency of Pre Trained BERT model and evaluate the same with Baseline Models like Logistic Regression and Support Vector Machines. The results show BERT model outperforming other baseline models with an increase in accuracy of 7 - 10 %. This report also sheds light on why BERT performs better than other Baseline models and discuss on classification mechanism of all the models.

Sentiment Analysis across Independent Datasets Using Recurrent Neural Networks

The performance of a Naive Bayes classifier compared to a LSTM and bidirectional LSTM recurrent neural network is investigated in the context of predicting the sentiment of movie reviews. The effect of using training data that is independent from test data is tested and analysed. It is concluded that the Naive Bayes classifier outperformed the neural networks in the context of this task and that the accuracy of all classifiers is lower than ones in related work where training data and test data are not independent in the same way.

Sentiment Analysis on Pfizer Vaccine Tweets with Supervised Learning

With the year 2020 being overrun by a pandemic that has resulted in its consequences being collectively felt and endured everywhere in the world, it would seem befitting to end the last semester of 2020 with a topic very much related to this. As 2021 has sprung up, we come closer to a solution to the ubiquitously cumbersome virus, also known as COVID-19, chiefly in the form of vaccines. One of these is the so called ”Pfizer vaccine”, developed by the companies Pfizer, BioNTech and Fosun Pharma. So in this rather novel and unique worldwide situation, what are people’s attitudes towards this vaccine, and with the use of supervised learning, can we learn to identify these sentiments? In this project this question was examined through a set of collected tweets and various text mining techniques. The results showed that the decision tree, AdaBoost, and neural network classifiers all performed the best, with accuracies surpassing 70%.

Comparison of Transformer and non-Transformer Models for a Text Classification Task

Abstract‚ÄîThe Natural Language Processing (NLP) field is changing with the introduction of the Transformer in 2017. The Transformer has replaced many Machine Learning (ML) methods achieving state-of-the-art results. However, in some particular cases such complex and computationally expensive models achieve results very similar to other, simpler, models. From this regards, 3 different ML approaches are trained on a dataset that provides information of Headlines and short descriptions of news articles from the Huffington Post and classifies them between different categories, such as Politics or Entertainment, among others. In this scenario non-Transformer models achieve similar results to the Transformer models for large amounts of data while needing much less training time. On the other hand, Transformer models achieve substantially better results for more difficult tasks and small amounts of data. All this results are thoroughly analysed, providing a clear image of when to use a Transformer-based model.

Exploring semantic qualities of a novel information retrieval system

In this paper a novel information retrieval system is presented. The system aims to find the best suited researcher given a free text query. The data used in the system consists of scientific papers. Two query expansion methods are implemented, which adds semantic capabilities to the system. The performance of the methods are evaluated by exploring the effect they have on vector space. The results show that when considering the closest document to the query vector, the expansion methods make the information retrieval system perform worse. However, when the average distance to a set number relevant documents are considered both methods show promise although the results can be seen as inconclusive.

Temporal Sentiment Analysis – Trends in Reviews over Time

Number or reviews a product can get on the internet in an international stage is vast, which is why this has to be automatically checked and interpreted. This report presents a pipeline containing text digestion, sentiment classification and visualization of the reviews over time. Looking at games or other software the reviews can change depending on updates, new content or bugs introduced to the users. To validate the pipeline the author predicts how the reviews change over time for a game using prior knowledge and patch notes. The results shows that common machine learning models can be used with simple text features to make accurate predictions on the data.

Machine learning models used are Multinomial Naive Bayes, Linear Support Vector Machines, Logtistic Regression, K-Nearest Neighbors and Random Subsampling. For the text features only Term Frequency tested with unigrams. Further the visualization from the validation data resembles the prediction made before viewing and analysing the validation data.

Machine Translation – Greek to English

Machine Translation is a very interesting topic that is rising during the last decades. It refers to the procedure of translating text from one language to another, which is done completely by the computer. There are a lot of approaches for this topic. However, this project will focus on a specific architecture of Neural Networks, the Transformer. The Transformers experienced a great growth the recent years and they are proven to outperform, both in performance and in training time, all of the traditional Neural Network architectures that were being used so far. They are based entirely only on the attention mechanism. In this project we will focus on the 'Greek to English' translation. We will try to figure how the performance of a simple Transformer is affected by the size of the data by using the BLEU score for evaluation and also observing the translation of a specific sentence throughout the process.

The trade-off between model complexity and model accuracy

Sentiment analysis remains one of the key problems that has seen an extensive application of Natural Language Processing (NLP). Since the proposal of Transformers architecture in 2017 for NLP, NLP advocates have been using it heavily to solve different NLP tasks. The application of these sophisticated models has resulted in many simple NLP tasks taking a long time and more resources to accomplish. Sentiment classification is one of the essential tasks in text classification and NLP and is one that Transformers have a significant presence in among other text classification architectures. In this project, two different models are tested against the same dataset, and the performance, as well as the time and resources for those models were measured. The Bidirectional Encoder Representations from Transformers (BERT) is used in this project as a Transformer-based machine learning algorithm, and the Multinomial Naive Bayes(MNB) classifier is used as a simple, yet powerful, machine learning classification algorithm. BERT unsurprisingly performed better than the MNB classifier. Nonetheless, the resources and time it took BERT model for fine-tuning parameters, training, and testing were incomparably higher than that of MNB.

Sentiment Analysis on Tweets to Analyze the Acceptance of the COVID-19 Vaccine

In recent years there has been a growing interest in sentiment analysis, and various state-of-the-art models have been developed. Thus, it is of interest to evaluate these in real-world applications. Two different classifiers are implemented and trained using the labeled Sentiment140 Twitter dataset. The two classifiers are; a Bidirectional Long Short Term Memory classifier and a Multinomial Naive Bayes classifier. The performance of two classifiers is evaluated and compared to several baseline classifiers. The feature extractions of the two classifiers, in terms of sentiment analysis on tweets, are also further explored. Finally, a Bidirectional Long Short Term Memory classifier, achieving an accuracy of 85%, is used to classify self-crawled tweets regarding the vaccine against COVID-19. It is concluded that 59% of the tweets about the vaccine are positive, while 41% are negative, giving a rough estimate of the global acceptance of the vaccine. The results confirm the numbers of a global survey on the acceptance of the COVID-19 vaccine.

The Evolution of Lyrics in German Rap

German rap is a relatively new music genre that has grown quickly in the presence of online streaming. In this work we build a LDA topic model to find five prevalent topics in German rap lyrics. A focus of the work lays on tackling the problems of preprocessing German song texts. We present a timeline that shows how the lyrical landscape in rap has evolved during the last 30 years. It has moved from mainstream music towards battle rap in the early 2000s. In the last decade trap and street rap are taking over the scene. We furthermore show how Sido's change from being a gangsta rapper to being a loving father reflects in his lyrics. Finally, we cluster the artists with help of a 2D visualization of all rappers' lyrical preferences. It shows how almost all rappers of the early years are clustered together, whereas today there appears to be more diversity in the lyrics.

Hotel Review Prediction using Multinomial Naïve Bayes and LSTM

The aim of this paper is to compare two different classifiers for predicting hotel reviews. The data set used is retrieved from Kaggle.com, and contains 20491 hotel reviews with each a correlated ranking within the interval 1-5, where 5 is the most positive. Furthermore, under sampling is performed because the waist majority of the reviews in the original data set are positive. The classification models compared in the paper are Multinomial Naïve Bayes (MNB) and Long short-term memory (LSTM). The models receive an accuracy of 72% respectively 46% for the original data set. For the under sampled data, the MNB receives a minor improvement in macro precision, while the LSTM decreases in performance.

Classification with Part-of-speech Tagging and Dependencies

This project tries to estimate how much of an author's characteristics is made up of his/hers choice of words compared to the author's grammatical features. The project evaluates this by taking three forms of the input: the words themselves, the part-of-speech tags of the words, and the dependency labels of the words. Each of these inputs is fed to a naive Bayes classifier for multinomial models and a convolutional neural network to then be compared internally and externally. The results point toward that there is not much extra unique information about the author in the word choice rather than in just the grammatical features. The difference between the inputs is greater in the naive Bayes classifier but in the convolutional neural network there is sometimes a difference of only 3%!

Improving text classifiers with data augmentation

The performance of supervised machine learning models is often limited by the amount of training data available. Data augmentation is the technique of increasing the size of the data set by copying and modifying existing data, or by generating synthetic data. This paper investigates how the accuracy of a text classifier can be increased by generating new text based on existing text. The dataset used was the titles and topics of trending Youtube videos in categories ”Gaming” and ”Film and Animation”. The classifier was a logistic regression model predicting the category of Youtube videos based on titles. LSTM generators were trained on titles of videos in the two categories, respectively, and on different vocabulary sizes, eliminating rare words. The classifier was trained on data from the dataset as well as an increasing number of generated titles. The classifier showed an increase in accuracy when the vocabulary size was larger than 20 words, and when the augmentation size was larger than 100 titles. The best accuracy increase was 5.8% with a vocabulary size of 200 words and an augmentation size of 900 titles.

Classifying fake news – An evaluation of different machine learning models

There are countless examples throughout history of how fake news have been used as a tool for influencing the public mind. Although in today's day and age, the easy access to information enabled by the internet, has made the presence of such information more prominent. Finding automated solutions tasked with detecting such news pieces has therefore become more relevant than ever. This paper investigates the classification accuracy of Long short-term memory (LSTM), Bidirectional LSTM, and Extreme Gradient Boosting (XGBoost), when tasked with classifying fake news for a large, labelled dataset. The dataset contains news articles which are labelled as either 0 or 1, indicating if a news article is reliable or not. The title of each article was used as input data, and was pre-processed using Porter's stemming algorithm, together with one-hot encoding. Hyperparameter tuning was conducted in order to optimize the classification accuracy of each model. The results indicate that the LSTM model yields the highest accuracy for fake news classification with the given dataset, even though XGBoost was tuned with a larger search-space.

Analysis of Disaster Tweets using Logistic Regression and BERT

Considering the impact social media has created during recent times, it has become important for personnel within fields such as crisis management to accurately identify the messages present on social media. This report focuses on the classification of tweets from social media website Twitter saying whether a tweet is about a disaster or not. In particular, the emphasis is given on fine-tuning of a pre-trained BERT model and its performance is compared with a baseline Logistic Regression model when trained on both balanced and unbalanced datasets. Experiments performed within this project provide evidence of the robustness of BERT as it outperforms Logistic Regression classifier giving a better accuracy by approximately 7%. The importance of other evaluation measures such as precision and recall are also discussed.

Comparing the performances of dieffrent transformer architectures in text: BERT, GPT2 and XLNET

These days, there has been a tremendous increase in the generated text by different webpages, sensors and other sources. It is very difficult to search and read for important data in a particular topic from huge text retrieved from several sources. In addition to this, there is a lot of data redundancy in the retrieved text. The most efficient way to deal with this data is by summarizing the large amounts of text into small sizes. Selecting suitable model for the task of text summarization is important to get better summaries. In this project, some of the pretrained transformer language models which are used for text summarization are compared. Some of current state-of-the-art transformer technologies like BERT, GPT2, and XLNet are used for comparing the performances. The evaluation of the models is done by using ROUGE metric. It evaluates the summary generated automatically and gives Precision (P-score), Recall (R-score) and F1 score. From the results obtained, it can be observed that the overall performance of the models is almost same with very little deviation which can be said through the F1 scores. In terms of precision, BERT model performed better than other two models. In terms of recall, the models GPT2 and XLNet performed better than BERT model. The accuracy of these models can be further improved by fine-tuning them with the domain specific corpus.

Performance comparison for lyrics classification in Traditional and Deep learning methods

Text classification is an important part of Natural Language Processing(NLP). Various models, including traditional machine learning methods and new deep learning methods, have been discovered to help people speed up development in NLP both in research and industry. In this report, such models were applied to music's genre classification based on lyrics and generate prediction results. We analyze performance based on accuracy, model complexity, fitting time. The result showed random forest had the best performance in accuracy, around 63%, but not in fitting time and model complexity, SVM balanced accuracy, and fitting time. Deep learning methods didn't reach an expected outcome, accuracy reached around 50%. The analysis has its limitation due to limited computational power and time.

Identification of Fake News

In natural language processing, fake news detection is a critical but difficult investigation to be done as natural language is not always presented in a fixed form with fixed topic. In this project, different bag-of-words embeddings with two different vectorizers, count vectorizer and tf-idf vectorizer, will be used to test the performance of three machine-learning models on its ability to correctly classify fake news as fake news. Support Vector Machine classifier, Naïve Bayes classifier, and Decision tree classifier are trained separately by using titles of the news and contents of the news. The performance of the models will be tested on titles and contents of the unseen news. The models will be assessed with three criteria accuracy, precision, and f1-score. Python module sklearn is mainly used in this project.

Fake news identification using text classifier

In recent years deceptive news articles are becoming dangerous prospects for online users, and with most of the news being published as online articles, it is becoming difficult to track the origin and validate information based on the source. Using fake news for political and economic gain is a rising trend in online articles. Presently, two ways are dominating in the detection of fake news. The first approach is to use human analysts as fact-checkers and is limited to analyst's knowledge to detect misinformation, the second is to use machine learning-based natural learning processing to detect fake news. In this project, three machine learning based models were compared using Kaggle's Fake News dataset and evaluated against their performance to distinguish the articles. The Bidirectional Encoder Representations from Transformers (BERT) is used in this project as a Transformer based machine learning algorithm, Logistic Regression based Classifier is used as a baseline model and paired with Random forest as Ensemble learning method. Logistic Regression based classifier is selected as the final model for its low requirement of training time and computation resources with a higher recall score.

Semantic Similarity on Quora Question pair DataSet.

The project explores the task of semantic analysis by looking at Quora Question pair DataSet. We conducted intense cleaning of data and various idea for feature extraction and used different type of deep learning models with different word embedding. Our final finding was LSTM with attention layer using Glove word embedding is a best model in term of accuracy or loss and computational power. We achieved maximum of 84% accuracy on simple Bi-LSTM model.

Impact of different text vectorizations and embeddings in sarcasm identification

Converting text into numerical representations is a key part of Natural Language Processing (NLP). Vectorization methods can vary in complexity, and some capture linguistic features that others do not, such as contextual differences of words. Sarcasm identification in text is a highly context-dependent task where positive words can have negative connotations and vice-versa. This report focused on the comparison of different word vectorization techniques, namely Term Frequency - Inverse Document Frequency (TF-IDF), pre-trained Continuous-Bag-of-Words (CBOW) and distilled Bidirectional Encoder Representations from Transformers (distilBERT) embeddings and their sarcasm classification accuracy using a Linear Support Vector Classifier (L-SVC) against a random choice baseline. The results show that sarcasm classification benefits from context-sensitive embeddings: The baseline, TF-IDF, CBOW reached, respectively, 50%, 79% and 81% accuracy on test data. Meanwhile, distilBERT embeddings achieved 86% accuracy. The analysis had limitations as embeddings were not fine-tuned for the task at hand.

Taste of wine analysis and wine classification in different regions

This study is to explore how the features and characters are described among different regions. And the next step is to create a model to classifiy wine region bases on the wine reviews. The project used TF-IDF vectorizer and Count vectorizer of 130k samples from the top 10 regions to analysis adjectives and nouns in unigram dn bigrams that are represented the characteries of the region. Statistical models as Navie Bayes and Logistics Regression, as well as nueral network model as Multilayer Preceptron and Convolution Neural Networds, are used to classify region from the tasting reviews. The result indicated that the unigram model has a higher performance among all the classifiers to classifiy the region, and also the insights of wine features are extracted by comparing unigram and bigram models.

Query expansion via Word Embedding

In this generation of the internet, a vast amount of data and news are available on the internet, which has a sort of effect on our life, our decisions, the books that we read and TV's talks and shows that we are interested in. This work presents a pipeline containing text analysis of Ted's talks titles in lights of the timeline of the year's event and visualization of similarity measurement using NLP techniques over a particular year. These include text pre-processing, word embedding techniques, vector similarity. The results explain that simple machine learning principles can be used besides text features to make a decision based on the data. The machine learning models used are Word Embeddings, PCA, K-Nearest Neighbors and features vector. For the text features, only Frequency-based Embedding Using Count Vectorized. This report and code produced can be found in the project repository on GitHub.

Multi class classification using a recurrent neural network

In conclusion, one can see that LSTM-RNN is a valid method to use for multi-class classification of text data and reaches and accuracy of 58%, but the approach used in this report may not be the best implementation of it. A LSTM-RNN network model also seems to have an edge over a Naive Bayes classifier as it seems to be able to discern reviews where the contexts of the words are more important, which usually is in the middle of the rating scale. Also, when compared to similar work it seems like the pre-processing and embedding steps are important to achieve high accuracy as those are the steps that differ the most between this report and that of the compared reports. Although, a comparison of different pre-processing and word embedding techniques would have been beneficial to do in this report as well and is advised for future similar studies. In addition, the number of epochs when training the model seems to have a large effect on the final accuracy of the model. When run too many times it seems to lead to overfitting on the training data, and thereby also a lower accuracy for the trained model when used on the test data and on real future data. There also seems to be no set number of epochs for a model that aims to do multi-class classification and is instead different for each model so tuning for this is required.

Sentiment analysis on podcasts reviews using fast CPU algorithms

With the quick development of deep learning, computer scientists and engineers are paying less attention to the efficiency of the code, the energy used, the carbon footprint and the margination (of people who cannot afford powerful costly computers to run their algorithms) generated. The modern GPT-3 has as many as 175 billion parameters to train, making it prohibitive for most of the population. In this project, a sentiment analysis task on podcast reviews is done by using simple methods (for the baseline), two open-source algorithms and pretrained models. The algorithms used were fasttext developed by Facebook research and Light GBM combined with GOSS and Dart developed by Microsoft. The pretrained models were pretrained on Amazon and Yelp data. The objective is to find the algorithm that works faster and scores better predictions.

Generating text using LSTM GAN in PyTorch

The aim of the project is to generate text using generative adversarial networks (GANs), where both components are recurrent/LSTM neural networks. The training data set is a collection of internet comments from YouTube, Twitter and Kaggle also containing a (binary) negative sentiment indicator. The model output is evaluated by perplexity using the Markov chain model trained on the training data related to perplexity of test data. Sentiment input integration into LSTM GAN is discussed.

Generating Stack Overflow Titles Using an Encoder-Decoder Model

Text generation and text summarization is a field within Natural Language Processing. In this project, an encoder-decoder is used to generate titles to questions. The dataset used for this project is ‚Python Questions from Stack Overflow" available on Kaggle. To evaluate the generated titles the average ROUGE-1, ROUGE-2, and ROUGE-L metric were used. The model that produced the best title according to the metrics had an average F-Score of 0.16, 0.028, and 0.15 with regards to ROUGE-1, ROUGE-2, and ROUGE-L. Two different models were implemented, one using an encoder with an unidirectional GRU and one using an encoder with a bidirectional GRU. Three different dimension of the hidden dimension were evaluated for both models during the project. For future work, it would be interesting to investigate how different parameters such as the number of epochs, the size of the word embeddings, and vocabulary size affects the result.

Extractive text summarization using Text Rank Algorithm and text classification of news articles

In this paper, I propose a system to summarize a news article using one of the Extractive Summarization techniques and classify the summarized news article using the deep learning model. To train and test, the data set used are news articles obtained from British Broadcasting Corporation News (BBC News) [1] which have been annotated byfive classes: business, entertainment, politics, sports, and technology. The proposed framework uses the Text Rank algorithm to summarize news articles and a deep learning classification model is applied to classify them into classes mentioned above. Text summarizer combined with text classification helps many people such as academicians,politicians, business personnel, and etc. to read a large chunk of text in a shorter period of time. Moreover, they could even use the classes to filter the summarized news to ease their reading experience. In order to view the real working of these 2 models, news articles from both CNN News and BBC News channels are web scraped, summarized,classified and presented to the user through a web application.

[1] D. Greene and P. Cunningham, "Practical solutions to the problem of diagonal dominance in kernel documentclustering," in Proc. 23rd International Conference on Machine learning (ICML'06). ACM Press, 2006, pp. 377–384

Analysis of Multi Class Movie Review Rating Classification using Pre Trained BERT Model and Evaluating with Baseline Classification Techniques

State Language Models like BERT, GPT 2, XLNET, T5 are often fine tuned for basic down streaming classification tasks in NLP. In this report, I experiment the efficiency of Pre Trained BERT model and evaluate with Baseline Models like Logistic Regression and Support Vector Machines on the basis of the Prediction Accuracy. The report also shows awareness and understanding of relative work documented in scientific sources and the extra finidngs obtained through this experiment. The experiment shows BERT model outperforming other baseline models with an increase in accuracy of 7 - 10% and analyzes on why BERT performs better than other Baseline models and the effects of Preprocessing the raw text.

Multi Class Text Classification-Consumer Finance Complaints Dataset

This project performs a comparative study between traditional ma-chine learning models like Multinomial Naive Bayes, Multiclass Logistic Regression, Linear Support Vector Machine, Random Forests trained ona variety of features and the state-of-the-art BERT in a multi class classification task using the consumer finance complaints data. Features like raw count, TF-IDF, Word2Vec and GloVe were used to train the baseline models. Linear SVM with TF-IDF was found to be the best performing baseline achieving an accuracy of 0.65. The results also show that pre-trained BERT outperforms the baselines with an accuracy of 0.70. The report briefly discusses the challenges of text classification on data with unbalanced classes and how it is handled in this project.

Comparing Text Vectorization Techniques for Sentiment Analysis Task

Numerical representation of texts as vectors is necessary for text classification tasks, including sentiment analysis. This project compares the performance of different classifiers on IMDB review dataset based on four text vectorization methods: bag-of-words, bag-of-words with tf-idf weights, average of word vectors, and DistilBERT. The results showed that tf-idf vectorizer paired with logistic regression or linear support vector classifier has the best result with 88% accuracy on the test data.

Finding relevant board games through user defined description

This paper aims to automate the task of searching for a board game through the use of text mining. The goal is attempted using the two text mining-related approaches information retrieval and word embeddings. In addition, the optimization technique smooth inverse frequency (SIF) is applied to the word embeddings. These approaches are compared in terms of cosine similarity and general substance of the descriptions of the resulting board games. It is shown that the best result is achieved by adding weights based on the frequency of respective word to its word embeddings.

Text Classification of Heavily Imbalanced Data Using Similarity-Based Approaches

Solving classification problem in Text Mining has its unique ways where two documents are combined together and then classified. However, for the cases where comparing two documents and classifying for a relation(label to be specific) remained the same. Even though the classifiers are performing well with the combination of texts, the curiosity of solving the comparison based problems by a using similarity approaches is the intention of this project. In the Machine Learning world particularly while solving classification problems, it is important to have a good amount of data for each class for a Machine Learning model to give better predictions. But, Data extracted from various sources is improper most of the times and will not have enough number of rows per each class all the time. In spite of imbalance in data, there is a high chance that Machine Learning models fail to generate better predictions. One way to solve the problem is by removing few rows and balance the data. This might lead us to loose some valuable information. There are various other techniques like Downsampling, resampling, generating synthetic data [1] and so on. Most of these techniques lead to either loose data or generate new data. This paper addresses the problem of classifying data without any type of resampling. With text data, there are many other ways to be considered to extract required information and classify the text based on that. This paper is an attempt to solve the problem of imbalanced data and classify the text based on Similarity rather than depending on Machine Learning algorithms.

IMDB Review Sentiment Classification using Machine Learning Models

In this work, we leverage machine learning methods for document classification. Specifically, we investigate different approaches to classifying the sentiment of reviews written on IMDB, divided into three different sized datasets. We find that for small datasets, traditionally shallow artificial neural networks perform worse compared to deep neural network models. Additionally, we propose an architecture that featurizes input sentences with an LSTM and feeds them through a multi-layered perceptron that performs relatively well. The LSTM is the best performer on the small dataset with 82% accuracy and the large dataset 88.2% accuracy. However, it takes significantly longer to train compared to the other presented models.

Emotional classification of Donald J. Trump's tweets

It has been reported in the media that Donald J. Trump is influencing the course of various market indices with his tweets. This offers potential for recognizing signals and developing trading strategies. One strategy could be to classify the tweets into emotions and then examine their effects on the market. Using the GetOldTweets-python functions, Donald J. Trump's tweets were scraped and then classified into emotions using IBM Watson Natural Language Understanding. Financial market data from Yahoo Finance was collected and linked to the newly created dataset to perform analyses. In order to circumvent IBM's limited service, a classifier was created to eventually replace IBM's classification service. Different methods such as dummy classifier, Naïve Bayes, Logistic regression, Bi-LSTM, were tested. An optimized logistic regression method achieved an accuracy of 0.719 on the test data and was the best performing classifier. It turned out that the emotion fear had a positive influence on the S&P 500 and can be considered as a possible trading strategy.

Hotel Review Prediction using Multinomial Naïve Bayes and LSTM

This paper aims to compare two different classifiers for predicting hotel reviews. The data set used is retrieved from Kaggle.com, and contains 20491 hotel reviews with each a correlated ranking within the interval 1-5, where 5 is the most positive. Furthermore, undersampling is performed because the vast majority of the reviews in the original data set are positive. The classification models compared in this paper are Multinomial Naïve Bayes (MNB) and Long Short-Term Memory (LSTM). Both models receive an accuracy of approximately 60% for the original data set (59.13% respectively 60.08%), but LSTM distinguishes by a low performance on recall and F1. This is assumed to be because of overfitting and is solved by the undersampled data. For the undersampled data, both MNB and LSTM receive an improvement in macro precision, but both models receive a decreased accuracy. Moreover, the models accomplish a more equal score for the undersampled data and the ratings individually receive a more equal score.

Comparing LSTM approaches with other methods on text classification

In this paper a comparison is made between LSTM text classification and other methods: Naive Bayesian, Decision Tree and K Nearest Neighbors classifiers. The performance is measured on two datasets: Wine Reviews dataset and Women's Clothing E-Commerce Reviews dataset with three different scenarios. We tested LSTM methods with embedding layer, pre-trained GloVe embedding and word2vec embedding. The results show that LSTM methods achieve similar performance to other methods. However, LSTM methods do not significantly outperform other methods. For a dataset with short length text, the LSTM with embedding layer has the best performance.