Project abstracts

Emotion Detection

Sensing the emotions and intention behind the words a human communicates has been a challenge for AI agents in general. Towards the end of the Text Mining course, the concept of context was emphasized. That there is more to just knowing the definition of the word being used. Emotion depends on the placement of words, the sequence of a whole sentence, etc. In this project, using a Bidirectional LSTM layer with the necessary preprocessing, a model to detect emotion from the text was built and evaluated. Showing a promising accuracy of 72% with other metrics evaluated in more detail in the report.

Sentiment Analysis of Apple Customers' Reviews on Twitter Using BERT-Base and Naive Bayes Models

In this project, the tweets of Apple customers on Twitter have been investigated. These reviews include three various labels Positive, Negative, and Neutral. If we want to apply a deep learning model on a text data set, where the words and text are our input, the learning method cannot recognize the words, definition of those words, an emotional load of the word, the usual position of these words, and so on. Using the BERT-Base model which is a powerful pre-trained NLP model can help us to implement a deep learning model on a text data set which is the reviews of Apple customers in this project. Furthermore, to evaluate how the BERT-Base model is accurate for this data set we implemented a Naive Bayes model and compare the accuracy of the two models. Ultimately as the data is imbalanced an oversampling method namely SMOTE has been applied for improving the model evaluation and it has been concluded that the result achieved by the BERT-Base model has been reasonably acceptable than Naive Bayes classifier.

Multi-Class Classification of Philosophy Schools

This study investigates the possibility of classifying schools of philosophy using multinomial naive Bayes and LSTM. The data consists of 13 different schools of philosophy like feminism, communism, and rationalism, to name a few. Since philosophy is a complex subject, it is interesting to investigate if there is enough relationship between the words in a single sentence to classify the input since each data entry only contains a single sentence. Or if it only depends on the topics discussed in the text. The project uses multinomial naive Bayes as a baseline and investigates if upsampling and LSTMs can increase the accuracy. Data used in this study were retrieved from Kaggle and consisted of 360808 individual sentences from the different schools of philosophy. For preprocessing the input for to the baseline was vectorized using a Countvectorizer from scikit-learn, and the input for the LSTM was tokenized, truncated, and padded using TensorFlow. The results showed that the baseline and the LSTM had an accuracy of 75% and 76% respectively.

Music Genre Classification From Song Lyrics

Most of today's music streaming services are providing many tens of millions of songs to their users. To not overwhelm the listener are there often ways to sort these based on genre, which of course relies on this information being stored and defined for each entry in the music database. This project explores the possibility of automatically deducing the genre for a specific song using its lyrics and to what extent this prediction can be accurately ensured. A large data set with over 380000 songs, including their lyrics and corresponding music genre, has been utilized to train two models to see how they compare in solving this task. A Naive Bayes model as the baseline followed by the transformer-based BERT model were implemented and compared using four standard classification metrics. The results show how the latter performs significantly better, and it was therefore chosen to be further investigated to find how the classification of specific music genres differed using the same metrics.

Genre Analysis of Movie Synopsis by Topic Modeling

Within the entertainment industry, movies are a high grossing sector and a frequent leisure activity. Prior to watching a movie, it is often important for the consumer to know some information about it in order to choose a movie which they are interested in. The purpose of this study is to investigate whether movies can be clustered together to generate coherent topics. By utilizing topic modeling methods Latent Dirichlet Allocation and BERTopic on movie synopses, the results show that it is possible to create coherent topics, particularly exhibited by BERTopic. The results from BERTopic also highlight that the keywords from the generated topics can be mapped to one or multiple movie genres. A (semi-) supervised approach can also be used to guide the topics toward their respective genre, but the unsupervised method also generates coherent topics.

Bitcoin Tweets Text Classification on Buy and Sell Actions

This project performs a text classification task on tweets regarding Bitcoin. The goal was to classify every tweet in a day containing the word Bitcoin or BTC into the labels Buy, Sell, and Neutral, based on how the Bitcoin price would change the following day. Then to use all the predictions to predict the future price change of the cryptocurrency. The model used was a word embedding neural network that also utilized additional Twitter features (retweets, likes, and replies). The model outperformed all baselines (probability dummy, most frequent dummy, and a Naive Bayes model). However, the accuracy of the model was at 39

Influence of Various Pre-Trained Word Embeddings for Sentiment Analysis

Performing an accurate sentiment analysis is an important task from many points of view. Understanding people's needs as well as understanding public opinion regarding impor- tant topics lets to adjust the approach and strategy for solving certain problems. In order to perform such accurate sentiment analysis, a machine must be able to understand the examined opinions. One of the representations for the text are word embeddings. Those multi-dimensional vectors can be trained from scratch as well as found to be pre-trained by others. In this study an LSTM network fed with such word representations was set up and the influence of both randomly initialized as well as pre-trained embeddings in the sentiment analysis of Tweeter posts was examined. It was shown that pre-trained and fine-tuned word embeddings have an impact on the accuracy of the predictions outperforming the baseline. Furthermore, dimensions of those vectors were shown not to have an impact on the predictions when using pre-trained fine-tuned word embeddings, whereas not fine-tuned word embeddings could have an implicit impact on the analysis. To perform the study, famous GloVe and fastText word embeddings of various dimensions were used.

Trip Adviser Data Augmentation

In classification the data amount is important for the performance of the model. If one class is underrepresented in the data-set the model performance can be worse for that class. This project studies two augmentation techniques that balances an unbalanced data-set. The data-set is Trip Advisor reviews and ratings and an existing classifier that classifies rating based on reviews is used. The augmentation techniques were oversampling and synthetic text generation using a RNN. The augmentation techniques increased the classifiers accuracy from 58% to 62% and 61% respectively. It also increased the per-class-accuracy of the underrepresented classes.

Can Language Models Learn the Style of a Text Author?

The aim of this project is to evaluate if text generating language models are capable to learn the style of an author on which texts/speeches models were trained. To do this, the definition of written style is analyzed, defined the components. For text generation, n-gram, LSTM based, and GPT-2 simple version language models were considered. Style similarity was evaluated regarding quality using the M-BLEU4 score and style strength using two trained classifiers. For the first classifier naïve bayes, gradient boosting, SVM and logistic regression with word count vectorization input was considered. The highest accuracy was obtained with naïve bayes classifier = 0.92. The second classifier is BiLSTM based where inputs are parts of speech tags and their dependencies, classifier obtained an accuracy of 0.85. Using these metrics, language model generated texts were compared with original D. Trump rally speeches and other politician rally speeches using Mann-Whitney U test. With a significance of 0.05, none of the language models were capable to generate texts which would be similar to real D. Trump speeches in all three criteria. Thus, there is no evidence to say that in this project considered language models are capable to learn the style of the author.

Self-Training for Binary Sentiment Classification of Movie Reviews

The state-of-the-art machine learning methods for classification usually require a large number of high-quality labeled training data and the task of manually annotating observations can be time consuming. Semi-supervised learning is an approach that combines both labeled and unlabeled data during training, with the intention of improving the predictive performance. This paper aims to evaluate the effectiveness of one semi-supervised learning method, namely self-training, for binary sentiment classification of movie reviews from the IMDb data set. The supervised classifiers Naive Bayes, Stochastic Gradient Descent (with Logistic Regression as loss function) and Random Forest are considered, both individually as baselines as well as in combination with self-training. For each combination of classifiers, several experiments are performed with different n-gram techniques and different proportions of labeled training data. The main results show that self-training in combination with a supervised classifier did not outperform the corresponding supervised baseline, in terms of accuracy, F1-score, precision and recall. This study underlines the importance of applying self-training with care and being aware of its limitations.

An Empirical Study on Author Classification for Swedish Chronicles

I present, to my knowledge, the first empirical study in author classification of Swedish chronicles. A study made by Books on Demand (BoD) shows that one in every third swede wants to write a book. With the demand growing and more people becoming an author. This empirical study is trying to find the most optimal classifier for Swedish chronicles. In the paper, it investigates if a more advanced solution like BERT can perform better compared to more ”simple” classifiers. The ”simple” classifiers chosen for this study were Multinomial Naive Bayes, Multinomial Logistic Regression, and Random Forest. The baselines selected were a Dummy classifier with the strategy ”most frequent” and ”stratified”. The result showed that Multinomial Logistic Regression had an impressive 91% accuracy.

Methods for Document Embedding Creation for Long Documents Using BERT Embeddings

The purpose of this project was to investigate and evaluate the use of RNNs (Elman, LSTM & GRU) to create document embeddings for book descriptions that exceed the input length for BERT. As a baseline a method established by other researchers were used. The initial token embeddings came from BERT and were aggregated using the methods mentioned above to create document embeddings. These embeddings were then used to train a simple neural net for a classification task, using genres of books as classes. The result obtained is that, with enough training, the models trained using Elman RNN or GRU embeddings were comparable to the model trained on the baseline embeddings, while the model using LSTM embeddings was not comparable.

Distinguish True and Fake News Titles

For humans, detecting fake news is a hard task. Addressing the problem of fake news detection using machine learning is one solution. This study approached the problem by classifying news by titles to allow clickbait prevention using an ensemble approach. Single classifiers of the methods Support Vector Classifier (SVC), Multilayer Perceptron (MLP), and Random Forest Classifier (RFC) were investigated with different text representations. The representations were bag of words (BoW) or term frequency-inverse document frequency (tf-idf), and unigrams or uni-/bigrams. The single models were combined within and between the methods by soft and hard voting approaches. A baseline model of Naïve Bayes, BoW, and unigram was used, which showcased a test accuracy of 0.9474. In comparison, the top-performing individual model consisted of MLP with BoW and uni-/bigram with a test accuracy of 0.9659. The overall top-performing model was an ensembled model by soft voting, consisting of the top-performing models per method selected by test accuracy. The accuracy obtained was 0.9705, and the highest f1-scores for both the label classes was obtained by that model.

Text Summarization With LSA Approach on WikiHow Articles

This study tried to summarize articles from WikiHow by using Latent Semantic Analyse (LSA). The data used is composed of 500 ar- ticles extracted from the WikiHow dataset. LSA is an extractive method that pick the k most important sentences in the original text to generate a summary. This project eval- uates different methods to select sentences with different types of text representation. The evaluation method uses is ROUGE for Recall-Oriented Understudy for Gisting Eval- uation and the study presents three of them: ROUGE-L, ROUGE-1 and ROGUE-2. We found that LSA performs badly on small summaries and we made a hypothesis about the best select sentence method for WikiHow articles.

Improving Section Belongings for Swedish News Articles - Density-Based Clustering Over Time Periods

For a newspaper to manually decide news sections based on the content can both be cumbersome and labor-intensive, especially in an environment where there exists a lot of noise and constantly evolving topics. This project investigates the possibility to automatically define groups of articles that share similar topics based on Latent Semantic Analysis and Hierarchical Density-Based Spatial Clustering of Applications with Noise together with a sliding window approach that is applied on a weekly basis. With a visual inspection of clusters, both the number of clusters and activity in each predefined section can be seen to vary over time. When comparing the predefined sections to the generated clusters, logical relationships could be identified. Manual evaluation using cluster summaries shows that the clustering method is able to find news events such as the conflict between Gaza and Israel and Tour de ski.

Automated Question Generation Using Transfer Learning

This project aims to build a novel Question generation system from a given passage. Given an input context(passage) we extract meaningful answer choices and feed the model with context, answer pairs to generate questions. The project explores the limits of transfer learning application using large pre-trained transformer model archtectures such as BERT, T5, mT5 and BART. The dataset used in the project is Stanford question answering (SQuAD) dataset. Further, the project is extended with a focus on MCQ type question generation, in which the system automatically fetches multiple answer choices that closely and semantically relates to our original answer choice. The project has widespread applications in education industry for example in automatic question generation for tests such as reading comprehensions, in human-computer-interaction for Q&A assessment and also in secure user authentication when answering questions rather than using predefined set of questions that are easy to hack.

Emoji Prediction for Short Text Messages

Emojis are symbolic characters that are widely used in today's digital writing, for instance in chat messages and posts on social media. There are several thousand symbols available, and they are used in different contexts. This report presents an ensemble, multi-class, multi-label prediction model utilizing a TF-IDF vectorizer, linear regression, and multinomial NaïveBayes, which predicts suitable emojis sequences based on text data. The model is trained on several million posts from the social media platform Twitter (so called "Tweets"). The different model options achieve precision and recall between 8% to 16% where a stratified random dummy classifier scores around 2% and a majority class dummy classifier scores around 11%.

Unsupervised Document Classification Using Topic Models

Document classification is a widely used application of Natural Language Processing. One of the major problems when it comes to document classification is the availability of labels in the training data. This is usually a time-consuming task since one would have to go through the whole article and assign labels to each article. This project aims to build a Topic Model using LDA to understand the underlying Topics within those documents and try to come up with a single deterministic topic for each document by using a classifier. In other words, this project aims to see if the results of a topic model can somehow be used to assign labels to the training dataset. In this project, I present two approaches that can be used to assign labels to the training documents.

The ultimate goal of this project is to create labels that are not the best but can be used for further classification

Bow Wow as in the Artist or Two Separate Words? Examining Text Classi Ers Using Bag-of-Words and Word Embedding Representations

Classification of text documents is an important task that serves many purposes. To perform a text classification, the text often needs to be represented in a vector shape. These representations lead to significant information loss. In this report, I present the results from an examination of how the performance of text classification models is being influenced using two different representations of text. Namely, Bag-of-Words (BoW) and word embedding representations. I examine this by training three classifiers (Naïve Bayes and XGBoost using a BoW representation, meanwhile Multi-layer Perceptron classifiers using the pre-trained BERT embedding) to detect ”spoilers” in movie reviews from the IMDb. The findings of the performed experiments indicate that raw word embedding does not improve the performance of the text classifier, compared to the classifiers using BoW representations of text.

Performance Evaluation of Deep Contextuality on Genre Classification in Music Lyrics

The transformer architecture has made an entrance into the NLP scene by its ability to effec- tively perform transfer learning between tasks after being trained on a vast majority of data. We evaluate the classification performance of the BERT model against other baseline models without deep contextuality such as the GloVe representation of word embeddings on genre prediction in music lyrics. Similar to work on previous data sets, the BERT model outperforms the other mod- els. Music lyrics tend to be long sequences of text whereas the lyrics themselves tend to be very similar across different genres. Computational complexity as well weighted f1-scores is evaluated across the models.

Performance of Pretrained DistilBERT for Authorship Attribution

This project investigates the performance of DistilBERT, a pretrained lightweight variant of BERT. This is done by comparing DistilBERTs embedding ability to more traditional feature extraction tools such as Term frequency - inverse document frequency (Tfid) and bag of Words (BoW) representation. These are compared on authorship attribution, where the goal is to predict an author from a given tweet. Logistic regression and random forest are used to utilize these encodings and make prediction on who wrote what tweet. Additionally, to preprocess the original data such that the classes are evenly distributed, a comparison between different sampling methods such as oversampling and undersampling was made. In addition to this, the effect of keeping or removing retweets was also investigated in conjunction with the sampling methods. It was shown that oversampling outperforms the other methods and the removal of retweets offers increase in average accuracy if and only if sufficient data is available. Furthermore, undersampling was shown to remove too much data, leaving classifiers with too few training samples. DistilBERT was shown to outperform the other methods in many aspects, however, DistilBERT was not fine-tuned to the data and thus did not scale as well with training samples as Tfid and BoW.

Stock Price Prediction From Written Text

In this work, we investigate the usefulness of text analysis for stock price prediction. We use a forecast system that predicts if the stock price changes will go up, down, or stay the day after the release of 8-K documents. We show that our models are better than random guessing, meaning that relevant information can be extracted from 8-K financial reports. We also demonstrate the use of dimensionality reduction technics to be able to process huge corpus. In particular, we show that we can drastically reduce the number of dimensions, without losing much performance.

Predicting Songs Metadata From Lyrics

Music is a worldwide phenomenon, as it increases in size and more and more music has produced the task to classify different metadata such as genre and loudness increases with it. Therefore in this paper, we look at different metadata for songs and try to predict them based upon their lyrics. This prediction is done with the help of embedding the lyrics in a pretrained embedding and then trying to predict with both K-nearest neighbors and logistic regression. The results show that although the genre is the easiest to predict other metadata such as the acousticness of a song and the release date show promising signs.

Balance Algorithm Evaluation

The interest in using data mining and natural language processing to solve complex real-life problems is steadily increasing. One problem often encountered with real life data is imbalances in that data collected. Imbalanced data can often have a large impact on the result when training an algorithm by creating bias for the overrepresented category of data. This can lead to worse performing classifiers or even make them useless. To counter this problem different methods have been developed to help balance the data. The goal of the project was to evaluate different methods for balancing dataset for natural language processing to see which ones produced the best result. Three different methods are evaluated in this project, Random Oversampling, SMOTE and Random Undersampling. The evaluation of the algorithms was done by comparing the results of a Multinomial Logistic Regression sentiment analysis. It was found that SMOTE and Random oversampling produced the best result with the least draw backs.

Classifying Scientific Papers With Graph Convolutional Networks

The rate at which scientific papers are published is increasing rapidly. While this undoubtedly means that a lot of interesting research is being done, the amount of material to sift through also adds friction to the research process. Automating tasks that currently involve human labour, like organization and quality assessment, is therefore desirable. In this work, the use of text classification to determine a paper's field of study and peer review is considered. Specifically, the task of text classification is turned into a node classification problem. A graph of documents and words that models global word co-occurrences is constructed, and a graph convolutional network is used to learn embeddings jointly for both documents and words. Results show that the method can outperform a baseline naïve Bayes classifier for the two considered tasks. Furthermore, the model performs particularly well when the number of labelled samples is limited. The resulting embedding space is investigated and turns out to be able to provide further insights into the documents and words in the corpus.

Evaluating the Predictive Performance of Reddit Posts on American Stock Prices

The problem of predicting future stock prices is a popular research topic. Traditional approaches resort to time series analysis and quantitative modeling, but recent methods using text mining and natural language processing approaches have shown promising results as well, as they are able to leverage information stored in text that contains general sentiment and beliefs that the public holds towards stocks. This project approaches the issue of stock price prediction by formulating it as a binary text classification problem, where the stock either goes up or down, and uses sentiment analysis, document vectorization and metadata from Reddit posts to train a classifier to predict price movements. Furthermore, the project examines if a moving average sentiment score in the form of a polarity score on its own can predict stock prices. The experimental results indicate that simply using a moving average of the sentiment of Reddit posts related to stocks is not a strong predictor of future stock prices. However, a linear SVM classifier trained on features extracted from Reddit posts can achieve a test accuracy of 90% when trained and evaluated on data from the year of 2021, which is about 30% higher than the naïve majority classifier baseline.

Evaluating the Cost & Performance of Different Text Classification Methods Using the Large Movie Review Dataset

Humans are subjective creatures by nature and thus, one's opinion can have a significant impact on their behavior. Even more so, when it concerns the success of products and services. In the entertainment industry, movies are largely profitable products and so, film reviews can have a huge impact on the performance of a movie in the box office since they hold the power of influencing consumers. This project utilizes the Large Movie Review Dataset containing reviews from the IMDb website and evaluates the cost and performance of five text classification methods in terms of predictive accuracy and training time. By utilizing a mix of three baseline and two neural methods, the results show that it is possible to predict sentiment of movie reviews with high accuracy with the baseline methods (Logistic Regression, Naïve Bayes, Support Vector Machines). The neural models (BERT-base uncased, BiLSTM) used were also able to achieve high accuracies compared to each other, although they required significantly more training time and gave somewhat lower predictive performance than the baseline methods.

Comparing Different Methods for Predicting Movie Ratings From User Reviews

The Internet Movie Database (IMDb) ranks movies based on average ratings from user reviews. However, some reviews lack ratings. This project use transfer learning to fine-tune a pre-trained Bidirectional Transformers for Language Understanding (BERT) model to predict ratings from reviews. To do this a dataset was created by web scarping review/rating pairs from IMDb's website. A classification model was then implemented and the result was compared, in terms of accuracy, to related work by Adhikari et al. [1]. The classification model was able to correctly classify more than one-third of the reviews. This was worse than the accuracy presented by Adhikari et al. [1], which was expected due to hardware limitations and dataset differences. In order to better estimate the average rating of multiple reviews, two other models were implemented: an interval squeezed regression model and an ordinal classification model. Both models were implemented such that each review was associated with a class so that they could be used for review classification as well. These models were compared to the classification model by the Root Mean Squared Error (RMSE). The lowest RMSE was achieved with the regression model.

Predicting Warnings of Ao3 Fanworks by Analyzing Chapters or Additional Freeform Tags

AO3, short for Archive Of Our Own, is a nonprofit open-source repository for fanworks contributed by users. In order to prevent AO3 users to read content that is too sensitive for them, six different warnings were added to the website. Authors can choose them, accordingly to the story they wrote. Aim of this project is to predict the warnings given to fanworks, based either on their chapters or their additional freeform tags that the author added. Data set has been scraped from the AO3 website and preprocessed via TfidfTransformer. Since a work can have one or more warnings, it is a multi-label classification problem, and two multi-label classifiers (BinaryRelevance and Label PowerSet) are trained and tested using the scraped data set. These classifiers transform this multi-label classification problem into, respectively, multiple binary ones or a multi-class one. We will see that those classifiers are more successful using when analyzing additional freeform tags and excluding works where no warnings apply.

Stochastic Graph-Assisted Genre Classification

We investigate how Natural Language Process- ing (NLP) can be leveraged using network in- formation. Specifically, we perform genre clas- sification of books where some of those refer- ence similar books. We apply the standardized text mining processing techniques on the de- scriptions given for the books. For the trans- formation of the text, we use the naive count vectorizer as well as the more complex GloVe embedding. We show that a Graph Neural Network (GNN) can significantly outperform a Multi Layer Perceptron (MLP) which is used as baseline and solely operates on the text data. Furthermore, we present the limitations and implications of our work.

Predicting Videogame Recommendations

The video game industry has been growing rapidly with more and more games being produced every year. In an ever more crowded market, it can therefore be difficult for consumers to know which games are worth their time and money. On the distributing platform Steam, users can leave reviews and tag them as ”recommend” or ”not recommend” but this is not the case on other platforms. If such reviews could be automatically tagged it could potentially allow for aggregation of reviews from many sites to aid people in their choices. To achieve this a bidirectional LSTM as well as a DistilBERT model was trained on a data set of Steam video game reviews to predict if a review recommends a game or not. Compared to a Naive Bayes model with an accuracy of 0.82 as a baseline, both the LSTM and DistilBERT models achieved an accuracy of 0.83, which was lower than expected. One likely reason for this was a too simplistic preprocessing of the raw texts. A limited amount of computing resources also slowed the process of training DistilBERT models, such that only a few designs could be tried out.

Detecting Sarcasm on News Headlines

Online resources, for example, news, become overwhelming to us due to the rising importance of the internet in our daily lives. It leads to more and more online news providers using sarcastic headlines to draw people's attention. Headlines expressed in sarcastic ways can be misunderstood and cause trouble when the wrong facts get spread. Thus, detecting sarcasm automatically is in need now more than ever. In this project, LSTM, BERT, and Naive Bayes (baseline) models are built to identify sarcasm in news headlines using a labeled news headline dataset. All models are fine-tuned using grid-search and holdout/cross-validation to obtain the best sets of hyper-parameters. In addition, a hybrid neural network model from a related article is also reproduced and fine-tuned on the same dataset for comparison. As a result, both LSTM and BERT models show better performance, in terms of F1 score and accuracy, than the baseline model. The LSTM and the hybrid models are faster to train than the BERT model. The BERT model is easier to construct and takes fewer epochs to converge due to the advantage of transfer learning. BERT also outperforms LSTM and the hybrid model, with 86

Duplicating a Swedish Twitter User: Neural Networks in Action

For the past half decade, a common project among data scientists (at least on sites such as Medium and Towards Data Science) has been to create twitter bots, trained on a specific user and with the aim of emulation. Usually, the targets have been well known English speakers. In this project however, I've selected a Swedish politician of which the purpose is creating a model capable of generating tweets that shares the same sentiments as the original user. To create this bot, neural networks has been the method of choice. More specifically, the recurrent long short-term memory neural network has been trained on tweets scraped from Twitter using the Python module 'snscrape'. The final result did however not meet the expectations, and the model were unable to generate meaningful sentences. Although the results were somewhat anticlimactic, I'm left with ideas on how to improve the model based upon the results (and the knowledge gained) in this project.

Comparison of Models for Sentiment Classification of Movie Reviews

Sentiment analysis is a domain within natural language processing useful for investigating popular opinion with respect to an entity. The increasing computational power of modern computers have opened the door to deep learning which has inspired many new approaches to language modeling in recent years. One of the most powerful neural network architectures today is the self-attention-based Transformer - a key component in the pre-trained language model BERT. This study compares the performance of BERT and a naïve Bayes classifier at the task of predicting the sentiment of movie reviews using two benchmark datasets and puts the results into context of previous work. The naïve Bayes classifier was surprisingly powerful, but could not match the performance of BERT, which was more accurate in general and closer to the true sentiment in its misclassifications. However, a significant weakness of deep learning models is interpretability, which is an issue that lawmakers and researchers are becoming increasingly aware of. Choosing a simpler model that is slightly inferior to a complex model as BERT in terms of accuracy may therefore be preferable in certain contexts.

Tokenizations Impact on Classification

Tokenization is the first step in almost any NLP pipeline. This project investigates the question: ”How does using a subword tokenization method such as WordPiece impact the result compared to word-level tokenization?”. To explore this question classification is performed on the headlines of news articles and an ablative comparison between 3 differently sized models is done. The findings conclude that subword tokenization is detrimental for small models, but advantageous for sufficiently large models.

Multiple Class Recommended Reading Ageprediction for Children Stories Using a Recurrent Neuralnetwork

This paper covers one use of natural language processing and machine learning to classify the recommended reading age for children stories. The dataset used when training the model is taken from the top rated datasets on Kaggle. The data were pre-processed using lemmatization, stop word and non-alphabetic character removal. The algorithm is coded in python using external libraries such as Keras, Numpy, Pandas and Tensorflow. The network was successfully trained using a Recurrent Neural Network with a Gated Recurrent Unit on a fairly small dataset. Different combinations of parameters, optimizers and batch sizes were tested out to find the model that performed best in terms of accuracy.

Extractive Text Summarization of News Articles Through BERT Embedding Clustering

With an ever growing amount of textual data available, the need for information extraction and summary is increased. Regarding news media, many articles and TV segments exceed the length of the average human attention span and several companies are creating news summary applications for easy user consumption. This project explores the concept of Extractive Text Summarization, where content is extracted and compressed from the original document while not modified, and implements and evaluates a previously proposed system for this concept. The system, utilizing the deep learning model BERT and k-means clustering, is evaluated on the CNN/Daily Mail benchmark and the results are comparable to the 13th placed system for the challenge. While many limitations are considered, the project provides metrics for a method previously used, but that has never been metrically evaluated before, which is regarded as a contribution.

Investigating the Possibility of Using a K-Meansclustering Model as an Unsupervised Classifier

In this paper, an investigation is made on how well a K-means clustering model can be used as an unsupervised classifier. This is done by looking at how well a K-means clustering model can do in terms of predicting and grouping together documents of similar type, essentially making the K-means model classify documents. A central part of the project includes studying the impact which different word vectorizers have on the clustering, as well as documenting the effect of preprocessing the text before clustering. Another part of the project investigates if an imbalanced dataset impacts the clustering result. Lastly, the predictions of a Naive Bayes classifier, as well as a Logistic Regression, are used to compare the clustering results. The results show that both the type of word vectorizer as well as the choice of preprocessing the text canaffect the clustering results, and thus the possibilities of using a K-means clustering model as an unsupervised classifier.

Classifying TV Characters by Analyzing TV Show Transcript

This project investigates how practices used for author- ship attribution can be used to classify characters from the TV show Game of Thrones, giving a classifier any sen- tence that character has spoken during the series. The performance of the three following classifiers are com- pared, Multinomial Naive Bayes, logistic regression, and a deep neural model. Using both term frequency and term frequency-inverse document frequency are examined for all models. How the models perform using different sizes for the training dataset and by evaluating the models of sen- tences of different lengths are also investigated. Overall the Naive Bayes model performed the best having an accuracy of 0.324. The deep model outperformed Naive Bayes only when trained on a very small dataset. The relative perfor- mance of the classifiers did not change when varying the length of the sentences used to evaluate the classifiers.

A Simple Recurrent Neural Network for Game Score Prediction Using a Game Description

One of the most fascinating parts of Machine Learning is how a computer is able to find structure and patterns in data, such as texts, using only logic. By using recurrent neural networks a program could even learn what contextual information is important for a prediction. One interesting question that springs to mind then, is if it is possible to find patterns in data where humans might not expect there to be any. This study uses a simple long short-term memory model in order to predict video game score from a video game summary. The data set used includes all the video games (1995-2021) from Metacritic. The results of this study conclude that the task is indeed very difficult even for machines, and even predicting the score based on the score distribution results in a better accuracy.

Multiple Class Recommended Reading Age Prediction for Children Stories Using a Recurrent Neural Network

This paper covers one use of natural language processing and machine learning to classify the recommended reading age for children stories. The dataset used when training the model is taken from the top rated datasets on Kaggle. The data were pre-processed using lemmatization, stop word and non-alphabetic character removal. The algorithm is coded in python using external libraries such as Keras, Numpy, Pandas and Tensorflow. The network was successfully trained using a Recurrent Neural Network with a Gated Recurrent Unit on a fairly small dataset. Different combinations of parameters, optimizers and batch sizes were tested out to find the model that performed best in terms of accuracy.

Evaluating Siamese Encoding Strategies for Semantic Code Search

Semantic code search is a problem in text mining and natural language processing that intends to retrieve relevant samples of source code given a query expressed in natural language. A common approach utilizing supervised learning employs separate encoders for sequences of source code and corresponding natural language. The objective is for aligned pairs to have encodings with high cosine similarity. This work explores utilizing shared weights across both encoders in a Siamese architecture compared to the non-shared approach – effectively halving the parameter count – and furthermore compares the effectiveness of a complex language model pre-trained on source code semantics, CodeBERT, with a general-purpose compact language model, DistilBERT. The models are trained and evaluated using mean reciprocal rank (MRR), training time, and the number of model parameters on the CodeSearchNet Python corpus, which uses in-source documentation as a proxy for natural language queries. While possibly tainted by inconsistencies in training hyperparameters across the models, the results indicate that DistilBERT can achieve MRR scores within five percentage points of CodeBERT where the latter incurs an additional factor ×1.8 increase in training time. Furthermore, the Siamese DistilBERT variant achieves MRR scores within one percentage point of its non-Siamese counterpart.

Neural Networks for Authorship Attribution in Victorian Era Literature

Authorship attribution is the task of identifying an author of a written text-piece. Most literature is signed, which implies that the author of a text is known. However, there are cases where the author of a text is unknown. One such situation could be unfinished work, another the result of misuse of citations or perhaps anonymous authors. The dataset used in this project is the Victorian Era dataset, with 1000 words per row and 45 different authors. This project was conducted to identify suitable methods for text classification to identify authors of textual works. A baseline classifier was implemented and compared to a neural network (NN) classifier and a long-short term memory (LSTM) classifier. A part-study was also conducted to analyze how these classifiers are affected by varying the number of authors present in the dataset.

Using a rather simple NN structure turned out to be the most effective when trying to identify authors. As an accuracy of 0.99918 was obtained, 3.142% and 0.568% better than the best baseline and LSTM classifier respectively. It was also found that the baseline classifier works best for small datasets while the LSTM classifier seems to benefit from larger ones.

Visualising Twitch.tv Emote Semantic Similarity Through Pre-Trained Language Models.

Online subcultures throughout the internet have, over time, developed unique variations of common languages. One such way in which these differences manifest themselves is through the use completely platform and community specific emotes. This work aims to produce a visual representation of how these sometimes obscure expressions relate to more widely spread ones through learnt word embeddings. Said embeddings are obtained by further pre-training a widely available language model on a wide variety of publicly available Twitch.tv chat logs

Comparison on Spam Email Classification With Different Methods

Imbalanced data can have a significant influence on learning system. There are two methods in transforming imbalanced data into a balanced one, oversampling and undersampling. To process text using machine learning or Neural Network models, text data need to be encoded into vectors of numerical values. There are two typical methods for text processing, Term frequency–inverse document frequency(tf-idf) and word embedding. In this project, four cutting-edge models are applied which are Logistic Regression, Support Vector Machine, Random Forest and TextRNN(LSTM) to explore an optimal combination of methods and models for spam email classification. The results exhibit that (1) Undersampling can be used instead of the imbalanced one to obtain an optimal classification performance. (2) Although both undersampling and preprocessing by removing information on original data can have equivalent performance based on the Support Vector Machine model, undersampling can save human time. (3) Many steps should be applied to transform the data into a specific form that can be used in TextRNN(LSTM). Although both TextRNN(LSTM) and Support Vector Machine can perform on the same level in small data, the simple model, support vector machine, can be an optimal selection.

Classification of Cuisine Based on Ingredients - A Comparison of Random Oversampling and SMOTE

There is a vast number of food recipes on the web, and many more are added each day. Users are cruise for new recipes to cook by cuisine, which makes it important for these recipes to be correctly categorized. Going through all these recipes manually would be very labour-intensive and time-consuming. One solution for this is predicting cuisines with the help of different classification methods. This project consists of three main parts: data processing, oversampling methods, and classifiers. The data was processed with Spacy's named entity recognizer to remove non-food related entities. The three different classifiers where Multinomial Naïve Bayes, Bernoulli Naïve Bayes, and Support Vector Machine. Furthermore, a comparison was made with the three different classifiers and the data balancing methods random oversampling and the Synthetic Minority Over-sampling Technique (SMOTE). The main results are that the model that performed the best was the Support Vector Machine model which got an approximately 87-87% accuracy for the balanced datasets with 8 cuisines and 72-73% accuracy for the big data set with 16 cuisines. As for the data balancing methods, they performed the same with similar accuracies, precision, recall, and F1-score.

Sentiment Analysis for Stock Price Prediction

This paper aims to investigate if sentiment analysis of Twitter posts can be used in order to predict stock market prices. Different sentence vectorization methods, classification models and datasets are explored in order to deduce if there is a correlation between Twitter sentiment and the future stock price for the mentioned company. Sarcasm and subtle nuances are recurrent in Twitter posts which tend to be weaknesses in many automated sentiment analysis methods. To try and mitigate this problem, *Bidirectional Encoder Representations from Transformers* (BERT) [**?**] sentence embeddings are looked into. A pipeline of BERT sentence embeddings and support vector classification is shown to achieve a high f1-score. By estimating the average sentiment of stocks at market opening time and comparing it to the following price change of that day, some indications of correlation is found.

News Classification System

This project illustrates the process of how text classification is carried out using various Machine Learning and Deep Learning Techniques. The main purpose of this project was to show that LSTM performs better than TfidfVectorizer for Text Classification. Data pre-processing has been performed using techniques like StopWord Removal and

Lemmatization. The text was then transformed to feature vector using TfidfVectorizer, upon that Classification Algorithms were used to classify text. Accuracy was calculated to see how well the Decision Tree and Random Forest can predict unseen data. The results obtained are compared with each other to know which algorithm provides the highest accuracy. Decision Tree gave an accuracy of 61% and Random Forest gave result of 86% accuracy. And then a recurrent neural network model, LSTM was implemented. Among all the methods used, Long Short-Term Memory (LSTM) provided an accuracy of 92% which is better than the results obtained from the other two algorithms.

Empirical Study for Bot Recognition Using BERT and BiLSTMs

Pre-trained language models together with the fine-tuning practices have been experiencing a remarkable improvement at NLP tasks during the last years. These models are learnt from well-written text corpora that share characteristics of academical and literary grammar, syntax, and lexicon. Literary works, Wikipedia and News are a few examples of these. When it comes to Tweet Spambot Recognition, we must deal with a big drawback: the difficulty of leveraging the prior knowledge of language models when applied on data that belong to an especially different domain. Social media text, such as Twitter tweets, features many peculiarities (own memes, emojis, hashtags, "at" symbols, misspellings, grammar incorrectness, etc.). In order to explore how these language models generalize on these web data distributions, we explored some Deep Learning language model architectures to research how they perform at the tweet Spambot Classification task. We used a fine-tuning approach to test the classification performance of BERT, Bi-LSTMs, and other classic Machine Learning techniques at several tweet data datasets. Our results show several experiments that yielded satisfactory performance. Hence, further research is needed to exploit the text-level classification against the traditional meta-data techniques that spend a lot of time and computational resources.

Evaluating Predictive Power of SEC 8-K Forms on Stock Prices

In the US, publicly listed companies must file a report called SEC 8-K (or Form 8-K) to inform their investors of important events in the company, such as bankruptcy or change of management. These forms follow a clear structure and may contain information that affects the stock in the days following the report, which has prompted researchers to explore processing the information using NLP in order to forecast stock prices. In this project, a deep neural network called LSTM was developed and used with GloVe word embeddings to predict UP or DOWN signals for stock prices. Importantly, only textual and no financial features were used for prediction. A unigram model with a random forest classifier was used as baseline. Despite attempts at tuning the LSTM model, it achieved 50.05% accuracy on test data, indicating that it was not able to find a predictive signal in the textual data. The unigram model achieved 52.98% accuracy, lending some weight to the usefulness of the textual information, but the majority class classifier still achieved the highest accuracy of 53.68%.

Bayesian Evaluation of Text Classification Models

When evaluating text classification models, we want to be certain about the performance of a model as well as its superiority over another. In the area of text classification it has become a norm to apply Null Hypothesis Significance Test(NHST) to statistically state, and compare classifier performance. But, a frequentist approach has its own limitations and fallacies. In this report, we reflect on limitations posed by NHST. We also implement a novel Bayesian approach for evaluating text-classification models. We use a benchmark dataset and create several shallow models consisting of sparse and dense features, and also an attention-based model for comparison. We empirically demonstrate the difference between the two evaluation approaches.

Multi-Class Classification of Clinical Specialities From Medical Transcriptions

Clinical text-based data has been explored to classify diseases based on symptoms and to classify medical documents. Supervised and unsupervised learning methods have been used in these works and researchers have published results supporting both methods. The recorded clinical transcriptions typically belong to a multitude of medical specialities and are stored in a large database. Manual parsing of such database is tedious and, in some cases, not feasible. However, an automated system utilizing a well-trained classification model can perform the job with minimal human interaction. The primary focus of this project is to classify medical transcriptions using four well known supervised classification techniques, namely Multinomial Naive Bayes (MultinomialNB), Linear Support Vector Classifier (LinearSVC), Logistic Regression and Random Forest. Text data was normalized and tokenized by Term Frequency-Inverse Document Frequency (TF-IDF) matrix. To improve the classification work, four re-sampling methods (RandomOverSampler, SMOTE, RandomUnderSampler, SMOTENN) and a feature reduction method, truncatedSVD was used before applying classification methods. Among the used four models, LinearSVC achieved highest accuracy (81%) after applying RandomOverSampler for the medical transcriptions data.

Tokenizations Impact on Classifcation

This project investigates the impact of subword tokenization on small rnn models when used on headline data to classify article category. The hypothesis is that the data will contain many rare words such as names of individuals and that subword tokenization can mitigate the out-of-vocabulary issues faced by word-level tokenization. The results does not bear this out conclusively.

Sentiment Classification in the Review of Genshin Impact

Nowadays, video games are a vital part of the entertainment industry, and mobile games play an essential role. Recently, a mobile game named Genshin Impact gets more and more popular and has achieved many international awards. People talk about this game on Twitter and Reddit. Therefore, we want to do the sentiment analysis on the comments of Twitter and Reddit. However, we do not have labeled data of comments in Twitter and Reddit, so in this paper, we would like to use the reviews of Genshin Impact from the Google play store as data to build a sentiment classifier for classifying the comments on Twitter and Reddit. Finally, we found that the classifier based on the BERT has the best performance among the models in this paper.

Comparison on Spam Email Classification With Different Methods

Imbalanced data can have a significant influence on learning system. There are two methods in transforming imbalanced data into a balanced one, oversampling and undersampling. To process text using machine learning or Neural Network models, text data need to be encoded into vectors of numerical values. There are two typical methods for text processing, Term frequency–inverse document frequency(tf- idf) and word embedding. In this project, four cutting-edge models are applied which are Logistic Regression, Support Vector Machine, Random Forest and TextRNN(LSTM) to explore an optimal combination of methods and models for spam email classification. The results reveal that undersampling data transformed from imbalanced data has the highest accuracy. Based on undersampling data, Support Vector Machine with tf-idf and TextRNN(LSTM) with word embedding both have the highest classification accuracy reaching 96%. According to the model complexity, the optimal spam email classification model is Support Vector Machine with tf-idf.

Sentiment Classification on Coronavirus Tweets

Since the end of 2019, coronavirus disease(COVID-19) has changed the world and the sentiment of the public.Twitter, as a large-scale public opinion platform, plays an important role to reflect the influence of the pandemic on the public sentiment. Based on the labelled tweets data of users' sentiment on the coronavirus disease, some supervised models(non-deep and deep models), can be trained to automatically label or predict the users' sentiment of new tweets. In the report,it compares the performances of these models. Also,there are some discussions on data processing and model extension.

Classifying Emotion in Text - A Comparison Between Different Datasets

Detecting emotion in texts is an important step towards making machines better at interacting with humans. There exist many text datasets that are labeled with emotions. When assessing the performance of models it is interesting to investigate how well the models generalize across datasets. In this report, three different datasets, each consisting of emotion-labeled tweets were investigated. First naive Bayes, logistic regression and BERT were trained and tested separately on each dataset. Then the different models were tested on the two other datasets, that they were not trained on. Using BERT tended to get higher F1-scores than logistic regression and naive Bayes. Naive Bayes performed badly on the less frequent labels unless the training data were balanced. The F1-scores decreased significantly when switching to training on one dataset and testing on a different dataset. The author suggests that more work is needed to investigate the quality of the labeling processes and/or the properties of the distributions of text in the datasets.

How Generative Are Generative Language Models?

In this project it is investigated how 'generative' generative models actually are and how they perform on extractive question answering tasks. This is done by comparing T5 models to a base RoBERTa model on the SQuAD data set. The results from this project indicate that generative models can be a competitive alternative for extractive QA tasks. The results also indicate that the output from generative models are similar to the output from extractive models, which indicates that generative models are perhaps not as generative as the name suggests and that the output is still largely based on the context provided. The main concerns for the viability of generative models for extractive question answering are that the diminishing returns for higher performance might be higher than their extractive counterparts, as well as the performance decreasing the longer answers are expected to be. In short, generative language models do not seem to be as generative as their name suggests, with similar output as their extractive counterparts.

Stochastic Graph-Assisted Genre Classification

We investigate how Natural Language Processing (NLP) can be leveraged using network information. Specifically, we perform genre classification of books where some of these reference similar books. Those relations resemble the graph data in our work while the text represents the node features in this network. We apply the standardized text mining processing techniques on the descriptions given for the books. For the transformation of the text, we use the naive count vectorizer as well as the more complex GloVe embedding. We show that a Graph Neural Network (GNN) can significantly outperform a Multi Layer Perceptron (MLP) which is used as baseline and solely operates on the text data. Furthermore, we present the limitations and implications of our work.

A Comparative Evaluation on BERT to Traditional Classifiers on Sentiment Analysis Classification

Increasingly more people can express their thoughts and opinions on the internet. We can take these thoughts and opinions and analyze them and from the results further improvements. This project aims to compare the Bidirectional Encoder Representations from Transformers (BERT) classifier to the more traditional machine learning classifiers. The classifiers will be evaluated based on accuracy and training time on a dataset of amazon reviews. The traditional classifiers that will be used in this project are Naïve Bayes, Random Forest, and Support Vector Machine (SVM). Since Naïve Bayes is the more common classifier, it will be used as the baseline. The reviews are divided into two datasets, one for training and testing. They are cleaned and preprocessed before the classification process begins. The results show that BERT is the more accurate classifier but also the slowest. Naïve Bayes got the worst accuracy but the fastest training time, Random Forest was the middle ground with high accuracy and low training time. SVM got the highest accuracy out of the traditional classifiers, but also got the highest training time.

Trip Adviser Data Augmentation

In classification the data amount is important for the performance of the model. If one class is underrepresented in the data-set the model performance can be worse for that class. This project studies two augmentation techniques that balances an unbalanced data-set. The data-set is Trip Advisor reviews and ratings and an existing classifier that classifies rating based on reviews is used. The augmentation techniques were oversampling and synthetic text generation using a RNN. The augmentation techniques increased the classifiers accuracy from 58% to 62% and 61% respectively. It also increased the per-class-accuracy of the underrepresented classes.

Predicting Fluctuations in Stock Price Using Sentiment Analysis on the Twitter Posts.

Factors such as social media posts and news articles influence the supply and demand of corporate stocks. This project aims to find the relationship between the sentiment of general public and the fluctuation in stock trend. It also aims to exploit this relationship to predict future stock fluctuations. Sentiment analysis is performed on the tweets related to Electronic Arts Inc. (EA) stocks. Sentiment score along with the lagged values of the stock variations are used as input to different machine learning models such as Linear model, decision tree model, random forest model, XGBoost model and support vector regressor model to learn the underlying relationship. Most accurate result is obtained with XGBoost model.

Inferring Correspondences Between Surnames and Social Status From Surname Co-Occurrence

Earlier vector embedding studies have focused on words and language use. In this paper I construct a vector embedding model for surnames from Wikipedia, and I then investigate whether analogous social roles can be found in a similar way to what previous studies have found for words in word embedding models. Out of the two main experiments that I conduct, the result of one shows strong support for social analogies and the result of the other disproves that those analogies are universally valid. The discrepancy is unusual enough, however, to warrant further study of a qualitative nature.

Identification of Spoiler in IMDB Movie Review

This paper presents the NLP (Natural Language Processing) approach to detecting spoilers in the IMDB review. Generally, these reviews reveal some information associated with the plot of a movie. An automated approach, filtering out such spoilers, would be ideal as manual labeling is impossible due to a large amount of content. To identify those reviews, we propose supervised machine learning models. So, we explored Bi-LSTM, XGBoost, Naive Bayes, and pre-trained GloVe to improve the accuracy in text classification. In addition to this, we used the Bag Of Words

(BOW), cosine similarity, and Term-Frequency and Inverse Document Frequency (TF-IDF) method to process the text vectors. The results shown from our models are satisfactory. Quantitative and qualitative results demonstrate the proposed method substantially outperforms the baseline model. Keywords: Bi-LSTM, GloVe, NLP, Spoiler, word2vec, Word embeddings, semantic approach.

Hotel Review Classification Using Imbalanced and Balanced Data

Choosing a good hotel for your stay is one of the most challenging steps of traveling, and if we had a machine learning model to tell us how well rated a hotel is, we could make an easy decision. In this project, three different classification algorithms will be evaluated; Decision trees, Gradient boosting and Multinomial Naive Bayes. They will be evaluated on a data set from Kaggle that contains around 20,000 hotel reviews and ratings on different hotels. Additionally, undersampling will be done on the training set to achieve a balanced dataset and to address the imbalance of the classes. The classifiers will then be evaluated on both the imbalanced (default) and balanced (undersampled) data sets. The classifiers accuracy f1-scores will be compared against each other and against a Dummy classifier baseline. The measurements used in this project will be the accuracy f1-score. The main results implied that the Gradient Boosting classifier is the most accurate classifier to use on this data set. It performed the best out of all other classifiers on both the balanced and imbalanced data. Overall, all classifiers had an improved accuracy score when the training set was imbalanced. It is discussed in the project that the reason behind this is the loss of information due to undersampling.

Identifying Short Term Stock Price Trends Through Financial News Articles

Predicting the stock market is a hard task with many different strategies available. In this paper the correlation between the sentiment of a company news article and the following stock price movements is analysed. The senti- ment of the articles is calculated using a pre- trained BERT model, fine-tuned on financial texts. The sentiment is then compared against the general trend of the stock price after the release of the article. The final results show some correlation between the sentiment of news articles and the calculated trendline, with the most informative result at short time intervals of when an article was released.

Bow Wow as in the Artist or Two Separate Words? Examining Text Classi Ers Using Bag-of-Words and Word-Embedding Representations

Classification of text documents is an important task that serves many purposes. To perform a text classification, the text often needs to be represented in a vector-shape. These representation leads to significant information loss. In this report, I present the results from an examination of how the performance of text classification models are being influenced using two different representations of text. Namely, *Bag-of-Words* (BoW) and *word embedding* representations. I examine this by training three classifiers (Naïve Bayes and XGBoost using a BoW representation, meanwhile Multilayer Perceptron classifiers using the pre-trained BERT embedding) to detect ”spoilers” in movie reviews from the IMDb. The findings of the performed experiments indicate that a raw word embedding does not improve the performance of the text classifier, compared to the classifiers using a BoW representations of text.

Sentiment Analysis for Stock Price Prediction

This study aims to investigate if sentiment analysis of Twitter posts can be used in order to predict stock market prices. Different sentence vectorization methods, classification models and datasets are explored in order to deduce if there is a correlation between Twitter sentiment and the future stock price for the company mentioned in the posts. Sarcasm and subtle nuances are recurrent in Twitter posts which tend to be weaknesses in many automated sentiment analysis methods. To try and mitigate this problem, Bidirectional Encoder Representations from Transformers (BERT) sentence encodings are looked into. A pipeline of BERT sentence encodings followed by a support-vector machine classifier is shown to achieve a high f1-score with respect to the explored datasets. By estimating the average sentiment of stocks in Twitter posts at market opening time and comparing it to the following price change of that day, some indications of correlation is found.

Stance Detection of News Articles for Fake News Detection Using Language Model BERT

The increasing amount of online news resources and increasing usage of social media has made it much easier to spread news articles and fact-checking them more difficult. This project investigated how text classification could help recognize the truthfulness of news articles. This was done by stance detection on a dataset from the Fake News challenge from 2017. This dataset consists of news articles and headlines and a classification for each pair of their relation (whether they agree or disagree with each other, discuss the same topic, or are totally unrelated). This is to help human factcheckers in finding fake news. The model that was used was the BERT transformer model and achieved F1-score of 73%, an accuracy of 92%, and a competition score of 10,278. This was superior to the 2017 winner that achieved a score of 9,556.5.