Project abstracts

732A92

Sentiment Analysis and Deceptive Opinion Spam for Hotel Reviews

Customer reviews are a very useful tool for the hotels in order to improve their facilities and increase their revenue. While sentiment analysis (opinion mining) of these reviews is a common application of NLP, not many papers have studied the detection of another type of opinion spam known as 'deceptive opinion spam'. In this project, we first perform a sentiment analysis using an SVM model for positive and negative reviews. The evaluation of the model is done in a different dataset (reviews from a different source). We also perform a deceptive opinion spam (i.e. classify the reviews as deceptive or truthful) starting with a model selection between gradient boosting, SVM and logistic regression. We evaluate the optimal model (logistic regression) and we compare its results with the paper 'Negative Deceptive Opinion Spam' by Ott et al. in order to comment whether or not we manage to increase the accuracy of the classifier. From the results presented, it can be said that logistic regression surpasses humans in terms of accuracy for deceptive opinion spam.

Grade: B

Knowledge Synthesizer

'Knowledge Synthesizer' is a tool that can simplify learning new and old topics. This tool creates a mind map of a given keyword and classifies other related topics by categorizing using topic modelling techniques such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF or NNMF). As a proof of concept, Wikipedia was used as a source of knowledge and the output from the Knowledge Synthesizer was evaluated by means of a survey where 18 students from the 2018-2020 Statistics and Machine Learning batch participated.

Grade: D

Spam Classification on SMS using Recurrent Neural Networks

Identifying spam emails has been a very common text mining and machine learning problem and it has been done using many different methodologies and techniques with varying levels of success whereas identifying spam messages on phones is a relatively much more recent problem. With the increased usage of mobile phones throughout the world, SMS (short messaging service) has become a more desired and approachable way of communication compared to emails. This has also lead to marketing and promotional campaigns being run via SMS which is many a time undesired by the user. In this paper, I have focused on studying the spam/ham classification using Recurrent Neural Network and the performance of this method. This paper does that by referring to research papers and blog articles. The data set that is used is the SMS spam collection data set provided by the UCI machine learning repository. In the end, I have concluded with my thoughts on the discussed models and further improvement opportunities in this field of short text identification/classification problem.

Grade: E

Predicting Oscars

This report how to classify movie reviews to predict if the movie at hand is going to win an Oscar or not. I explain the different algorithms used, where Logistic Regression with TF-IDF was the one giving the highest result, and accuracy of 70%. In this project a huge amount of time was spent on data cleaning where I reduced a dataset of about 8.5 million rows down to about one million rows.

Grade: D

Classifying Stock Price Changes Based on 8-K SEC Filings with Gradient Boosting Decision Trees

This text mining project makes use of filings from the U.S. Securities and Exchange Commission (SEC), in particular 8-K filings, to forecast short-term percentage changes in stock prices. 8-K filings have to be published by public companies in the U.S. for certain defined, business-relevant events. The forecasting problem is approached as a classification problem with five classes: large decrease, small decrease, no change, small increase and large increase (of the stock price before and after filing date of the 8-K filing). Instead of using existing data sets, a new, up-to-date data set is created, with data for the year 2018 and the first quarter of 2019.

Furthermore, this project makes a contribution by testing a new approach to this problem: gradient boosting decision trees (GBDT) with tf-idf bag-of-words. Previous research had found random forests to perform very well on this problem. Hence, the methodological choice for this project fell on GBDT as a more advanced ensemble model. Exclusively text data was used for training the models. The best model could achieve an additional 10 percentage points of accuracy on the test data compared to the majority classifier, which is comparable to previous findings from the literature.

Grade: A

Exploratory Analysis of Music Lyrics by Discovering Relationships between Songs Based on Genre and Gender using Clustering Mechanisms

This report examines the relationship between the lyrics of numerous songs of different genres using text mining techniques. Multiple measures are taken to ensure whether songs of similar context belong to the same genre category or if they belong to singers of the same gender. Text clustering provides the most efficient and accurate mechanisms to group text data based on similar contexts. Therefore, this project has employed the vectorization approach to compute the term frequencies of each document's lyrics to obtain a matrix of tf-idf weights. On obtaining this matrix, two clustering algorithms are applied – K-means clustering and DBSCAN clustering. A suitable comparison between the distance-based clustering approach and the density-based clustering approach has also been explored through evaluation by finding their respective rand index. The rand index obtained in both cases depicts the accuracy of the clusters and can be used to conclude whether the clusters are grouped based on the same genre types or if the artists that sung the respective songs are of the same gender.

Grade: D

Predicting Myers–Briggs Personality Type from Social Media Posts

Personality explains the characteristics of a human being which includes their actions, thoughts, belief etc. Studies have shown that most components of our personality remain unchanged throughout except few that adapt to changes and show variation. Understanding personality is helpful in many aspects for an individual, the most important being self-awareness which in turn can help an individual in making the right choices and improve decision-making skills. Various researches are being conducted in past years in identifying the personality type of people and it was proved that the text written by a person is not independent of his personality. A tremendous

increase in the use of social media in recent years gives ample personalized textual data. Two popular models for assessing personality type is Big5 and MBTI. This project focus on training supervised learning models to predict the MBTI personality of an individual by relying on social media posts. The models are evaluated by measuring accuracy and F1-score. The publicly available social media posts of individuals are fed as inputs to the classifier. The models discussed in this project predict binary output class of the four dimensions of MBTI type individually and then combine those to predict the MBTI type. Building an accurate model will have major implications in areas where knowledge of personality benefits such as workplace, business etc. An accurate model can also benefit people in knowing their personality type without going through lengthy test procedures.

Grade: D

Impact of Contextualised and Non-Contextualised Word Embeddings on Classification Performance

This project investigates the capabilities of contextualised word embeddings provided by DistilBERT as a bidirectional transformer encoder compared to non-contextualised word embeddings provided by Word2Vec. As DistilBERT is based on encoders only, a feed-forward network (FFN) and an Long Short-Term Memory (LSTM) are used as models for the downstream task of text classification. The dataset used, are Amazon reviews ranging from one to five stars and the performance of all four combinations of word embeddings and classifiers are being tested with the reviews. The accuracies on the test data set and the models corresponding training times are: The FFN with Word2Vec achieved an accuracy of 43.10 percent, training for 2.3 hours; the FFN with DistilBERT achieved an accuracy of 49.98 percent, training for 11.65 hours; the LSTM with Word2Vec achieved an accuracy of 50.49 percent, training for 6.35 hours; the LSTM with DistilBERT achieved an accuracy of 56.43 percent training for 12.1 hours. Concluding, contextualised word embeddings help for the task of text classification at the cost of a higher computational demand.

Grade: A

Evaluating Classification Algorithms in Classifying Genetic Mutation Based on Clinical Evidence

Interpreting and classifying the type of mutations based on text-based clinical evidence could be a time consuming and challenging job. However, text classification

algorithm can facilitate the process by automatically classifying given clinical evidence into corresponding mutation class. To achieve this goal, 4 different classification algorithms were tested and evaluated in this project. Finally one classification algorithm that returned most optimal performance was selected and discussed.

Grade: D

Conventional Machine Learning Models versus Convolutional Neural Networks in a Multi-Class Text Classification Problem

The importance of accurate text classification algorithms is straightforward, however a consensus on which Machine Learning model is superior in text classification settings remains absent. This paper aims to determine whether state-of-the art Convolutional Neural Networks (CNN) outperform conventional Machine Learning models. A wide range of models is proposed: Naive Bayes, Logistic Regression, Decision Tree, Random Forest, XGBoost and Convolutional Neural Networks. In order to identify a superior classification model, a comprehensive model evaluation is performed. Effects of different vectorizers, CountVectorizer and tf-idf vectorizer, together with balancing the training dataset is analyzed. A simple CNN architecture consisting of 9 layers is assessed, as well as a more complex design consisting of 24 layers. Results show that a more complex CNN does not improve results. Considering all models, Logistic Regression shows an outperformance when using a CountVectorizer on an unbalanced training dataset. Balancing the training data improves recall rates throughout multiple models, however, implementing the tf-idf vectorizer fails to raise performance rates. All models are trained and evaluated on a Coursera review dataset consisting of 107,018 reviews together with ratings ranging on a 1–5 scale.

Grade: B

Sentiment Classification with Word Embedings on Amazon Reviews

The ability to extract meaningful insights from text data has become very important nowadays, with more sophisticated algorithms and computational power we can use huge amounts of data to build powerful models in order to use them for any related problem consisting of text data. Through this project we are going to build a sentiment classifier using some of the most recent state of the art algorithms such as XgBoost and LSTMs that are currently widely used to perform similar tasks using the open source framework Tensorflow/Keras [21], XGBoost [22] and Scikit-learn [23].

Grade: E

Topic Modelling on United Nations General Assembly Speeches 1989

Aim of this project is to discover topics from United Nations General Assembly speeches from the year 1989. The data, obtained from Kaggle, is preprocessed using the standard text mining pipeline of tokenization, stop word removal and lemmatization. Term frequency–inverse document frequency is used in Latent Dirichlet Allocation model instead of the bag-of-words representation. Interestingly, the countries grouped by continents had showed diversity in concerns with topics ranging from the revolutions in Europe, Apartheid in South Africa, civil wars and dictatorships in South America, drug trafficking in Columbia and nuclear weapons. However, there were some overlaps in topics which implies a common goal.

Grade: E

Performing Sentiment Analysis of Twitter Data to Analyse People's Happiness

According to the World Happiness Report 2019, people from less developed regions show significantly lower happiness than people from highly developed regions. Using the UK and a large region in southern Africa, this theory is examined by comparing Twitter data from the two areas. Multiple sentiment classifiers (Naïve Bayes, Decision Tree and XGBoost) using different combinations of vectorizer, pre-processing procedure and hyperparameter setting are trained and tested using the labelled Sentiment140 data set. A Naïve Bayes classifier identified as optimal is then used to classify previously self-crawled tweets from both regions. To increase the chance of a correct classification, classified tweets that are not assigned a minimum probability of 75% being in the positive or negative class are removed from the analyses. Following this procedure, the tweets considered are expected to be 85% correctly classified. The classifications of the self-crawled tweets are used to obtain the relative frequency of positive and negative tweets within both regions for different years. This is extended by the identification of named entities within the classified tweets. This combined information is used to analyse the happiness of the people in the two regions. The results obtained do not confirm the extreme differences in people's happiness between less and highly developed regions as presented in the World Happiness Report.

Grade: A

Tweet Analysis Using Machine Learning Algorithms

The explosion of the reach of social media with cheap availability of internet connectivity has brought with it its advantages and disadvantages. In my native country, India

one of the menace of social media is the spread of fake news through social media especially Twitter. This is the motivation for the project and the model is trained on the tweets from FakeNewsNet which is collections of fake and true tweets on Gossip and Political tweets. I have tried different methods to compare their performance including Naive Bayes and neural networks. Various types of classifications were attemtped to analyse tweets, trained and tested on the FakeNewsNet dataset. Then using Twitter API, real tweets filtered from India is selected and the models is tested on them. The tweets were attempted to be classified as 1. Two class classification -Fake or True Tweet 2.Two class classification-Political or Gossip Tweet and 3.Four class classification -Political Fake(PF), Political True (PT), Gossip True (GT), Gossip Fake(GF). These there classifications are independent classifications and help us analysing a tweet into differnt types. The "Transfer of Knowledge" from a knowledge base which consisted of tweets from the US used to classify tweets from another region (India) was also studied.

Grade: E

Genre Classification with Lyrics

In the field of Natural Language Processing, there are many ways to represent a word numerically. In this project, we focus to compare some of state-of-the-art embedding technologies on a classification task. This paper implements a music genre classifier from lyrics by using DistilBERT embeddings as contextual embedding and GloVe, Word2Vec as uncontextual embeddings. We compare their performance in the sense of accuracy, F1-Score, computational complexity and training time on this specific task. We use a Long Short Term Memory classifier. As a result, we observe that GloVe outperforms the others. BERT is in the second place without fine-tuning, but it has the most complexity. GloVe achieves 39% after 35 minutes, BERT achieves 38% after 500 minutes and Word2Vec achieves 28% after 285 minutes of training.

Grade: A

SentimentalCrypto: Classification of Sentiment of the Cryptocurrency Market and Clustering analysis

Most Cryptocurrencies use Blockchain technology to record and distribute ledgers of transactions. The largest by market capitalisation are Bitcoin and Ethereum. There exist over 5000 cryptocurrencies, most of which are widely speculated and traded over a large number of exchanges online. This project aims to build a sentiment classification tool to establish the current sentiment of the cryptocurrency market

based on Twitter feed. Further, Clustering analysis is performed to better understand the Twitter corpus. 320K tweets that were segmented by emotions: anger, fear, greed, hateful, joy, sadness were used to train our classifiers after preprocessing and classification training was done using Multinomial Naive Bayes, Logistic Regression, Stochastic Gradient Classifier and method with highest accuracy were chosen for further use. Valence Aware Dictionary and sEntiment Reasoner (VADER) was used to generate a polarity score and clustering analysis was performed on Twitter corpus using K-Nearest Neighbors. Finally, the chosen classifier was applied on completely unseen data pulled from Twitter using the Tweepy API to classify market sentiment as on 7th March 2020. Our classifier showed that the market was largely filled with 'Greed' and 'Fear' with a small propotion of 'Hateful'. On 15th March 2020, 'fear' had increased and mixed emotions of 'sadness' and 'joy' were also present. This project of classification of the current market sentiment was performed during the marketwide sell-off with drop in Bitcoin's price from 9000$ to 4000$ amidst the Covid-19 pandemic where even global stock markets experienced major declines in sentiment.

Grade: C

Effect of Trump's Tweets on Oil Price

Donald Trump, the current president of the united states, has always been very active on social media. His explicit words have been on the top of the news many times. On the social media platform Twitter, his posts(known as tweets) are the subject of this project. The goal is to acknowledge any correlation between his tweets and the oil price. To do so I performed sentiment analysis on a data set of tweets, and gave a total point to his tweets per day, and tried to compare them to the variation of in oil price. I used three different methods of sentiment analysis (NRC, AFINN, and Bing). The results, however, did not show any promising correlation.

Grade: E

Sentiment Analysis of Amazon Product Reviews Using Machine Learning Approach

Nowadays companies are trying to focus more on the quality of products as well as providing a convenient way for customers to express their thoughts or opinions about their products. It helps other buyers and companies in improving a particular product on the basis of customer reviews. In this project, I scraped data from the Amazon website, which is 80,000 product reviews of different categories (Books, Medicine, Fitness Equipments, etc) to perform sentiment analysis. I have 2 class levels positive

reviews and negative reviews. To perform analysis, I have selected two supervised machine learning algorithms 'Naive Bayes' and 'Support Vector Machine' to check which algorithm performs better on Amazon product reviews data.

Grade: E

Topic Modeling to Highlight the Main Message of Jesus in the Holy Bible of Christians

Jesus is the central figure of the Holy Bible and Christian faith. This study used text mining natural language processing techniques for topic modeling to highlight the main message of Jesus in the Holy Bible of Christians. This study used theory from the Structural Topic Model, Latent Dirichlet Allocation and Zipf's law. The study concluded that the topics that highlight the main message of Jesus are 'love', 'father', 'son' and 'god' and that 'love' was overwhelmingly the most important topic. This study also concluded by showing that the main topics remain consistent for most of the language translations of the Christian Holy Bible.

Grade: E

News Recommendation System

In today's e-commerce industry, recommendation systems play a major role. Many people read the news online by visiting news websites based on their interests. But due to the availability of a lot of online information that has been published every day on different websites, makes it challenging for readers to find the relevant information which they require. Hence, I decided to analyze the textual 'CI&T Deskdrop' dataset which is available in Kaggle and build a content-based news recommendation system to recommend reader's interest-based news articles and save their valuable time. Various NLP (Natural Language Processing) techniques have been applied for preprocessing the textual data. The feature extractors such as tf-idf (term frequency-inverse document frequency) and count vectorizer are used to extract the numerical features from data. Then similarity function (cosine similarity) has been applied and top articles are recommended based on the similarity score from two different methods. Based on the top recommendations and recall'@'k evaluation metric, a good recommender system has been suggested.

Grade: D

Sentiment Analysis of Hotel Reviews – Performance Evaluation of Machine Learning Algorithms

Customer reviews on hotels are very important part of travel plan for people now a days. People prefer to book such hotels which have high number of positive reviews. There are different sources to find the reviews to get a better insight about the hotel's reputation. Thus it can be said that customer reviews plays an important part for business owners in order to improve their services. In this project, sentiment analysis is performed on the basis of user reviews using three different classifiers. The classifiers used in this project are 'Naive Bayes', 'Random Forest' and 'Support Vector Machine'. The performance of these algorithms are assessed on two different parameter settings. The reviews are classified as 'positive', 'negative' or 'average' labels.

Grade: E

The Study on Distinction for Authentic and Overseas Chinese Recipes

Chinese food is an important part of Chinese culture, which influenced many other cuisines during thousands of years. Simultaneously, overseas Chinese restaurants thrive and their dishes differ from traditional Chinese recipes gradually. Based on our studies on 482 Chinese recipes, it is hard to find the difference between authentic and overseas Chinese food via some unsupervised learning skills. However, the hidden specific combination among cooking methods and raw materials can be found by such simply supervised learning models as linear SVM and two-layer neural network, which can be used for the classification for authentic and over- seas Chinese recipes.

Grade: D

Exploratory Analysis of the Characters from the TV Show 'Friends' through Text Mining with Network Graphs and Sentiment Analysis

Friends is an American TV show that is now 26 years old yet all generations of people around the world have loved it. This project investigates the main characters of this most popular TV show 'Friends'. To find out if there exists a main character or a central character at the least, amongst the six friends on whom the show is based on. In order to contextualize this Network Graphs are used to understand the relation between the characters. Evaluation and comparison are done through centrality measures by using Closeness Centrality, Eigenvector Centrality, and Page Rank Algorithm scores. The data used for Network Graphs is the whole transcript of the show which was pre-processed before using it. To understand the emotions of the characters for

personal interests and gaining more knowledge Sentiment Analysis is performed (It is a minor part of the project). For Sentiment Analysis, IBM Watson's Natural Language Understanding (NLU) API is used from IBM Cloud Services and the data is preprocessed in a different manner using the NLTK library.

Grade: B

Subreddit Classification from Post Titles Using Sentence Embeddings Computed from BERT and GloVe

This project investigates the classification of Reddit posts based on their post titles to appropriate subreddits using word and sentence embeddings. Specifically, this project compares the classification performance achieved by using contextual word and sentence embeddings from BERT with non-contextual sentence embeddings aggregated as the sum of GloVe word embeddings. The classification task is performed on two different datasets - a coarse-grained dataset with 17 classes and a fine-grained dataset with 1431 classes. In order to improve the classification performance, a label smoothing approach is evaluated and fine tuning of the BERT model is performed. The sum of contextual word embeddings from BERT after fine-tuning with label smoothing is found to achieve the best classification performance on both datasets. On the coarse-grained dataset, this approach achieves a test accuracy of 82.04% and on the fine-grained dataset, this approach achieves a test accuracy of 40.51%. Pre-trained BERT embeddings perform better than GloVe embeddings on the simpler dataset but fine tuning is necessary for the BERT embeddings to perform better than GloVe embeddings on the challenging fine-grained dataset. Label smoothing is found to consistently improve classification performance, but only marginally.

Grade: A

Toxicity in Dota 2 Early, Mid and Late Game

Dota 2 is a multiplayer online battle arena (MOBA) game. It is one of the most popular and highly competitive computer games. A particular aspect of this online game is the toxicity and harassment that players experience while gaming. In this project, text extracted from the in-game chat are processed using the generative model Latent Dirichlet Allocation to summarize different time stages in the game. Afterwards, a toxicity score is assigned to each sum- mary of the stage by making use of the PerspectiveAPI developed by Jigsaw and Google as an unbiased measure of toxicity. These stages are usually called early, mid and late game. Each of these stages are characterized by having different behaviours of the players and as such, different

levels of toxicity. This project tested my personal belief/hypothesis that early and late game are more toxic than mid game. The results of this project however shows that mid game is more toxic than late game and that late game is more toxic than early game.

Grade: A

Comparison of Text Classification Methods Based on The Songs' Lyrics over Decades

In the report, we aim to compare the performance of five different text classifiers, including multinomial Naive Bayes, decision tree, logistic regression, support vector machines, and convolutional neural networks, based on the experiment of classification of songs' publication decades by their lyrics. We evaluate the performance of classification by using different statistics and statistical plots, including F1 score, accuracy, ROC curve, and area under curve statistic. Finally, we figure out the logistic regression classifier performs the best among the five classifiers when we hope to classify the song's publication decades with lyrics.

Grade: E

Comparing Performance of LDA and TF-IDF on the Five News Groups Dataset

In the current situation, the world is surrounded by news articles. There are hundreds of publications for News articles in America alone, and each publication has a different agenda for publishing news articles. As a part of my Text mining project, I chose to analyse the news articles from some of the American publications, to see if we can successfully identify the publication from the content of the news article. The topic modeling technique, Latent Dirichlet Allocation, and the numerical statistic for textual data, Term Frequency–Inverse Document Frequency, were used to generate features for the classifiers. The performance of these features was evaluated using two classifiers, Multinomial Logistic Regression and Support Vector Machine. The aim of this project was to compare the performance of the feature generating algorithms, Latent Dirichlet Allocation and the Term Frequency–Inverse Document Frequency, on the News Articles dataset. Experimental results show that the Term Frequency–Inverse Document Frequency gives a better overall accuracy with both the classifiers on the News articles dataset.

Grade: E

Movie Review Classification Analysis Using Supervised Machine Learning Techniques

Movie review is an important and useful tool as an information source for the audience whom willing to attend a specific movie. Recently movie-goers build their decision for which movie to watch by examining movie ratings along with review browsing at movie-related sites. In this project, we performed text classification analysis using supervised machine learning techniques, namely, linear support vector classifier SVC, random forest and extreme gradient boosting XGBoost for different movie reviews to classify into 'positive' and 'negative', and aimed to build a comparision system between the employed classifiers and find the optimal one for this type of problem. To provide an unbiased evaluation of the model, holdout method is employed. The overall accuracy was high for all classifiers, though LinearSVC had the highest accuracy with 92.1%, while random forest and XGBoost accuracies were 88.9% and 87.0%, respectively.

Grade: C

TDDE16

How Essential are Lyrics for Understanding Songs?

This project investigates how essential lyrics are for understanding songs. The analysis includes song genre classification and hit song similarity measurement with LDA topic models. From the classification problem, it was found that multinominal Naive Bayes gets the best results with an accuracy of 88%. The study of hit songs with Wasserstein similarity between topic distributions of songs could not find any strong correlations between documents. A comparison of vector representations of the songs with cosine similarity found that the documents with the highest topic similarity did, in fact, not also yield high similarity between songs.

Grade: 5

GPT-2 Model Evaluation

In this project, we reviewed the pre-trained GPT-2 model as well as a LSTM binary classified with pre-trained word vectors. The goal of the classifier was to evaluate whether a text is generated by the GPT-2 model or if the text is written by a human. The pre-trained GPT-2 model was used to generate training and test data from the Wikipedia's movie plots dataset. The computationally expensive methods for text

generation presented challenges. Consequently, the number of generated data was low and the LSTM classifier received 69% f1-score on the text data for the generated texts, and 55% f1-score for the original texts. However, the limitation of the GPT-2 model needs further investigation. We conclude that the LSTM model shows promising results even with a low number of data points.

Grade: 5

Neural Network Approaches to Sentiment Analysis on IMDB Movie Reviews

Previous work has shown that there are several methods and ways for efficient sentiment analysis on text data. With the drastic increase in both performance and results, but most importantly popularity of neural network solutions in the latest years it has been proven to work reliable on similar tasks to a comparable or better degree than older techniques. In this work we do sentiment analysis on IMDB movie reviews. We create a baseline of more simpler methods such as Naive Bayes and LogisticRegression, where many reach a accuracy of around 90%. We implement two different neural networks – a standard deep neural network and a recurrent neural network – and compare these to the baseline. The most important concept of this work is to share as much between the models as possible e.g. transforming and padding of data, as the comparison in this work is between the models and not towards the highest possible accuracy. The results shows that the DNN reaches comparable results to the baseline, without much fine tuning (as desired). The RNN does not achieve as good results and has extremely high computation time. Through experiments and research we take the reasonable expectation that higher accuracy and lower computation time could be achieved for both neural network if specifying the configuration to each individual method, which is left to future work.

Grade: 4

Generating Headlines for News Articles: A Summarization Approach

This work studies the task of automatically generating titles for a given text. A summarization approach is taken to develop a method that utilize TF-IDF scores, sentence extraction and parse trees. The method is evaluated on the Harvard Dataverse News Articles corpus, containing 3824 news articles. Four judges were presented 20 generated titles and asked to rate them based on grammar, conciseness, relevancy and overall fit. The results show that the presented method can generate perfect titles with a success rate of 20%. A new approach for finding candidate sentences utilizing word

similarity is also explored and evaluated and is shown to outperform the baseline approach.

Grade: 5

Implementing and Testing Topic Models based on Anchor Words

Recent work in the field of topic modeling present computationally efficient estimation algorithms for topic models based on a separability assumption. The assumption, known as the anchor word assumption, assumes that there exists a unique word for every topic which cannot be found in any other topic. The algorithms derived through this assumption have provable bound and have a computational complexity effectively independent of corpus size. This project aims to implement one such estimation technique in Python and evaluate it compared to a Latent Dirichlet allocation model estimated using Bayesian inference. The corpus used for evaluation are descriptions of all Netflix movies and TV-series published as of November 2019. Due to unfortunate circumstances however the reliability of the results produced are questionable and may or may not be valid.

Grade: 5

Classifying IMDB Rating Based on Movie Reviews

With streaming services enabling people to watch movies anywhere at any time, the movie industry is larger than ever before. IMDB is largest database of information related to movies, and may be the most famous for having a rating on every movie widely seen as the most trust-worthy indicator to how good a movie actually is. In this project I analyze reviews on over 1500 movies to se whether I can predict if a movie is rated above or below average solely based on the text from the reviews. The classification of the text is done using VADER and then the output from the VADER classification is used as input to a random forest model to make the final prediction. The results showed an accuracy of 0.62.

Grade: 3

Evaluating Text Classification Methods for Automatically Assigning CVSS Scores to Software Vulnerabilities

Determining the severity of a software vulnerability is an important step in the vulnerability reporting process, as with limited resources it is usually most desirable

to prioritize those that could lead to the most damage to a company or customers. The Common Vulnerability Scoring System (CVSS) provides a system for assigning numerical severity scores, ranging from 0 to 10, to vulnerabilities according to a set of criteria that are evaluated by human experts. This project explores the accuracy of using different text classification methods to assign CVSS scores automatically based on textual descriptions of the vulnerabilities, and compares the results to previous work on this subject. The text features that the classifiers deem most important are also analyzed, which gives an insight into what the classifiers base their predictions on. Through the use of Bidirectional Encoder Representations from Transformers (BERT), a current state-of-the-art technique for natural language processing, the correct CVSS score could be predicted to the nearest integer with 59% accuracy. Simpler methods such as logistic regression and Support Vector Machines (SVM) also performed well, with accuracies of 55% and 54% respectively.

Grade: 5

Evaluation of Fake News Machine Learning Classifiers

Fake news is becoming an increasingly popular topic. The objective of fake news is to misinform readers, which is more likely to occur if the fake news are similar to normal news articles. Fake news classification can be addressed with machine learning, but it is important to find out which classifiers perform the best. There are many different methods, parameter and settings that can be used. With little knowledge about the data or about different classifiers it is difficult to know which to choose. This paper provides guidelines for different steps that can be done to distinguish the good models from the bad and could be applied to practically any text classification problem. It is shown that the classification performance is improved with the addition of auxiliary features and suitable vectorization techniques for different features. Out of the tested classifiers, Support Vector Machines with default settings and a Multinomial Naive Bayes classifier with sentiment are the best.

Grade: 5

Sentiment Analysis on Movie Reviews

In this project random forest is investigated to see if it is fit to be used for sentiment analysis, and how it performs compared to Support Vector Machine (SVM) and naive Bayes. IMDb reviews are used to train a model which should be able to classify reviews as positive or negative. TF-IDF features are extracted from a document set of 75,000 reviews. Random grid search is used to optimize the hyperparameters of the

model. Final random forest model achieves 86.5% accuracy on a class balanced test set. It does beat naive Bayes but not SVM, which scores 89.4% accuracy.

Grade: 3

Automatic Music Genre Classification Using Lyrics and Neural Networks

Today there are many different methods of automatic music genre classification. The best methods incorporates both audio and lyrical data to classify the genre of different songs. In this report, different neural networks are trained on only lyrical data and compared to find which performs the best when classifying five genres. The genres used in this report were: pop, rock, hip-hop/rap, country and R&B/Soul. A fully connected neural network (NN) was used as base line together with the the Naïve Bayes classifier. Addition to them, a recurrent neural network (RNN) and a convolutional neural network (CNN) were compared and it was shown that the CNN performed the best with an average accuracy of 85%. The networks were also evaluated on precision, recall and f1-score for each genre. Out of the five genres, Country and R&B/Soul were the easiest to classify while pop and rock were the hardest.

Grade: 4

An Explorative Study in the Taste Descriptions of Cheap and Expensive Red Wine

The goal of this study was to explore differences in how cheap and expensive red wine is described. The study used a TF-IDF vectorizer to find adjectives, nouns, bigrams, and trigrams that were specific to describing cheap and expensive wines respectively. Using a data set with taste descriptions provided by Systembolaget, it was found that cheap wines are often described with words related to fruits and berries, whereas expensive wines are described with more abstract and sophisticated terms. The results also indicated that there is no direct correlation between the price and the expert rating of a wine.

Grade: 4

Investigating if People are Defined by Their Actions

There is a famous quote suggesting that people is defined by their actions. In this paper we investigate that claim by extracting actions performed by three famous persons from a dataset of news articles. Using a SVM and a feedforward neural

network we attempt to classify which person did what action. The results indicate that the connection between action and person is higher than random. The best results was achieved by combining an action together with its sentence context.

Grade: 5

Multiclass Classification of Python Questions

This report outlines the development and evaluation of three different text classification architectures, or systems, used to classify Python questions from Stack Overflow as being tagged as one of 20 different tags. The systems employ different vectorization techniques, namely tf-idf and word embeddings, for representing textual features. Furthermore, the systems have a varying level of complexity, two systems employ simple logistic regression, while one system employs a more complex, albeit still simple, convolutional neural network (CNN) on top of logistic regression. The evaluation of the systems show that the CNN system achieves the best performance compared to the other systems, as well as compared to a simple baseline, mainly by capturing larger contexts within the texts. Furthermore, the results show that each system, even the very simple ones, can achieve good performance with very little hyperparameter tuning. The results also show that ambiguity within the texts are a serious inhibitor when it comes to text classification, even in a very limited domain.

Grade: 5

Judging Books by Their Cover

The aim of this project was to train a classifier that tags a book with the appropriate genres based on the description usually found on the back of the cover. Since a book can belong to multiple genres the amount of available labels had to be reduced. This was done by clustering the available genres according to word similarity using k-means and spaCy's vector representation of words. The clusters where then used in the classification which yielded greatly improved results.

Grade: 3

Question Answering Machine

In this project, an attempt to build a machine that finds an answer to a given question is made. For this, the Stanford Question Answering Dataset was used to train and evaluate the machine. The machine itself consists of different steps where words are

translated to vectors in a high dimensional vector space and after that compared to one another in order to determine what words in the context that are most likely to be the answer to the question. The machine aims to mimic a Bi-Directional Attention Flow network, by using steps that are simplified but that still achieves a similar prediction. The final model achieved a f1 score of 57.4%.

Grade: 4

Comparison Between Naive Bayes and SVM for Classification of Song Lyrics Sections

Song lyrics is almost always divided into different sections. These sections is then put together in some sort of arrangement which result in a final song. For example, one famous song structure is the following: verse, chorus, verse, chorus, bridge, chorus. This gives us tree different sections (verse, chorus and bridge), ordered in some type of way which result in a final song. This project will investigate if a Naive Bayes classifier and Support Vector Machine can be used to classify these sections. The lyrical data is related to the pop genre and was collected from Genius.com API using the python client LyricsGenius.

Grade: 3