Ideas for theses

DS1. Genre classification of songs

Devise a method that maps an input song into one or many genres. For example, a song X can belong to genres A, B, and C. Or, a song X can belong 80% to genre A, 75% to genre B, and 68% to genre C.

As the main novelty, in particular investigate the perception of genres between listeners. For example, the user Mary can annotate that song X is of genres A and B while Mohammad may annotate song X as being of genres B and C. The research questions can be:

- what is the genre variance?
- what are the characteristics of songs that have little genre variance (i.e. most of the users annotate them with the same genres) vs those who have a lot of variance?
- what are the characteristics of users that do annotations in line with the mainstream and those who are outliers?
- can a user-based model for genre classification be built (i.e. a trained model that assignes the genre labels based on the listener)?

Methodology:

- Investigate existing work
- Compile dataset
  - Existing
  - User study
- Feature Engineering
- Build ML pipeline
- Evaluate

Features:

- audio-based
- lyrics-based

Target

- genre

References

- Markus Schedl Genre-related stuff
- https://www.kaggle.com/c/mlp2016-7-msd-genre
- http://millionsongdataset.com/
- http://cs229.stanford.edu/proj2019spr/report/3.pdf

DS2. Genre Classification of Movies

Devise a method that maps an input movie into one or many genres. For example, a movie X can belong to genres A, B, and C. Or, a movie X can belong 80% to genre A, 75% to genre B, and 68% to genre C.

As the main novelty, in particular investigate the perception of genres between listeners. For example, the user Mary can annotate that movie X is of genres A and B while Mohammad may annotate the movie X as being of genres B and C. The research questions can be:

- what is the genre variance?
- what are the characteristics of movies that have little genre variance (i.e. most of the users annotate them with the same genres) vs those who have a lot of variance?
- what are the characteristics of users that do annotations in line with the mainstream and those who are outliers?
- can a user-based model for genre classification be built (i.e. a trained model that assigns the genre labels based on the viewer)?

Methodology:

- Investigate existing work
- Compile dataset
    - Existing
    - User study
- Feature Engineering
- Build ML pipeline
- Evaluate

Features:

- audio-based
- video-based
- poster-based
- trailer-based
- subtitles

Target

- genre

References

- http://cs229.stanford.edu/proj2019spr/report/9.pdf
- https://pdfs.semanticscholar.org/b3c5/a370c4e32ea7a095cf6b41441f0f4fe0d6f9.pdf
- https://ieeexplore.ieee.org/document/7727207
- https://www.analyticsvidhya.com/blog/2019/04/predicting-movie-genres-nlp-multi-label-classification/
- https://link.springer.com/chapter/10.1007/978-3-319-09879-1_32

DS3 Explaining AI outcomes

Often AI approaches (e.g. machine learning models) generate outputs (predictions, recommendations etc) for which the user does not understand how they were generated. For example, the user may wonder why a specific set of items were recommended. A bank client would like to understand why she was denied getting the loan. Ai algorithms often rely on hidden features (e.g. latent features in matrix factorization, deep layers in neural networks) that can't provide the required explanation.

Pick a topic of your choice (e.g. recommendations of touristic destinations), find a dataset, set up an algorithm and devise a method for explaining the results given by the algorithm so that the end user can understand.

Example:

- https://research.tue.nl/en/studentTheses/understanding-the-latent-features-of-matrix-factorization-algorit

DS4 Counteract bias in datasets

AI algorithms (e.g. machine learning models) generate outputs (predictions, recommendations etc) that are biased. For example, only popular songs get recommended, keeping non-popular artists little chance to gain popularity. In banking, people from certain cohorts get their loan applications denied more often than others. These biases shape both our personal decisions and the ones made by policy makers. The biases can be rooted in the dataset used for training the model or the model itself.

Pick a domain of your choice, analyse the biases (the cause), propose methods to counteract on the biases and pick an appropriate evaluation metric.

DS5 Sonification of other modalities

A fun way to play with data is to visualize it through audio signals. For example, the scenery seen through the window of a train during a ride can be mapped to music.

Pick an input signal of your choice and devise a method to sonify it. Examples:

- http://www.cp.jku.at/projects/soundtracks/
- http://cs229.stanford.edu/proj2019spr/report/1.pdf

DS6 Reciprocal Recommendations

Recommender systems are usually framed as "given user X recommend a set of items that maximize the utility for the user X". However, in certain scenarios, the item can also be an active user that has its own preferences. For example, in dating applications, recommendations must be done in such a way that both users' utility is maximized. Similar examples are recommendations of professor-student pairs for mentorship.

Picks a domain where reciprocity is required and develop a recommender system.

DS7 Recommendations for items

Recommender systems are usually framed as "given user X recommend a set of items that maximize the utility for the user X". However, in certain scenarios we would like something like "for item Y recommend a set of users whose utility would be maximized by item Y". For example, if a company is selling a product, they would like to know who are the most probable customers so they can focus their marketing campaign.

Pick a domain where the item is the central point of recommendations and develop a recommender system.

DS8 Modeling Personalized Perceptual Distances

Different Users perceive item similarity/distance in different ways. This is driven by expertise, cultural background etc. Applications that rely on similarity metrics (e.g. collaborative filtering recommendation techniques) may hence yield inaccurate results if the user-specific perception of similarity is not taken into account.

Pick a domain where there is a variance in perceived similarities (e.g. music) and devise a similarity method trained on user feedback. If possible, use the personalized distance in an application (e.g. recommender system).

DS8 Framework for Scraping Social Media Traces

Data science application, especially in the domain of computational social science, often rely on the data that users leave behind them when they interact in social media sites. These data can be useful for pure research, user feedback, marketing etc. There are two important aspects in collecting such data: (i) the engineering aspect, where an appropriate software needs to be developed to scrap the data and (ii) the ethical aspect of collecting and storing such data.

Develop a software framework for collecting data from several social media platfroms (e.g. Twitter, Facebook, Instagram, Amazon reviews etc.). Discuss the ethical implications, user consent, gamification of the data collection process etc.

DS9 Detection of Audio Elements

Given a dataset of audio recordings develop a model that is able to detect one or more characteristics or events within the audio: instrument detection, onset detection, beat detection etc.

Example:

- http://cs229.stanford.edu/proj2019spr/report/6.pdf

DS10 Detection of hate speech from text

Define hate speech characteristics that can be extracted from an essay (essay, product review, blog post, social media post etc.).

Build (find, augment) a dataset. Feature engineering, Machine learning workflow. Evaluation.

References:

- Presentation.pdf
- http://cs229.stanford.edu/proj2019spr/report/69.pdf
- http://cs229.stanford.edu/proj2019spr/report/79.pdf

DS11 Detection of fake news from text

Define hate speech. Identify content characteristics that can be extracted from an essay (essay, product review, blog post, social media post etc.).

Build (find, augment) a dataset. Feature engineering, Machine learning workflow. Evaluation.

References:

- https://data-flair.training/blogs/advanced-python-project-detecting-fake-news/
- Presentation.pdf
- http://cs229.stanford.edu/proj2019spr/report/69.pdf
- http://cs229.stanford.edu/proj2019spr/report/79.pdf

DS12 Detection of emotions/sentiment from text

Identify content characteristics that can be extracted from an essay (essay, product review, blog post, social media post etc.): emotion, sentiment (towards an object),

Build (find, augment) a dataset. Feature engineering, Machine learning workflow. Evaluation.

References:

- Presentation.pdf
- http://cs229.stanford.edu/proj2019spr/report/69.pdf
- http://cs229.stanford.edu/proj2019spr/report/79.pdf

DS13 Identification of gullible users

Can a machine accurately predict whether a social media user is gullible to fake news/scams/manipulations? Define gullible users. What are the digital traces left by gullible users that distinguish them from non-gullible?

Build annotated dataset. Feature engineering. Machine learning workflow. Evaluation.

References

- https://www.researchgate.net/publication/332813387_Identification_of_credulous_users_on_Twitter
- https://www.orau.org/impact/stories/research/can-a-machine-accurately-predict-whether-a-social-media-user-is-gullible-to-fake-news.html
- https://www.technewsworld.com/story/60520.html
- https://dl.acm.org/doi/10.1145/3292522.3326055

DS14 Identifying Political Orientation of Users

Can political orientation of people be identified from social media traces? Which social media behaviour distinguishes people belonging to different political options. Similar work has been successfully done in the past in the case of American voters, where they have a simple uniform bi-class distribution (Kosinski et al).

Apply the state of the art techniques to voters from a European environment with multiple classes (parties) that are non-uniformly distributed.

References:

- Kosinski et al, https://www.pnas.org/content/110/15/5802

DS15 Detecting Drivers Personalities

People's personality has several models one of the most known is the Five-factor model (FFM). The factors of the FFM can be inferred from user's social media traces (see authors, such as Golbeck, Tkalcic, Ferwerda, Kosinski). For drivers (car drivers) one could use the FFM but there are models more specific to the driving scenario, such as the multi-dimensional driving style inventory (MDSI). The user model can be inferred using features derived from social media or from the driving itself (e.g. through a mobile app that reads the car parameters through OBD2).

Collect a dataset, annotate the users with their personalities, machine learning workflow.

DS 15.1 Drowsiness detection

https://data-flair.training/blogs/python-project-driver-drowsiness-detection-system/

DS16 Detecting Insurance-expensive Drivers

From the perspective of car insurance, some drivers are more expensive, i.e. cause more damage tha the insurance need to pay. Define the labels according to the insurance perspective. Collect a dataset and devise a machine learning workflow for the detection of expensive drivers from features, such as social media, OBD2 signals etc.

DS17 Predicting Stock Values From Social Media

There is some evidence that social media activities and stock values are correlated. Collect a dataset and train a model that scraps social media info (e.g. twitter) and predicts the value of the stocks.

Examples:

- http://cs229.stanford.edu/proj2019spr/report/31.pdf
- https://www.sciencedirect.com/science/article/pii/S1877050917326194
- https://www.springer.com/gp/book/9783658095079

DS18 Multi-stakeholder Recommender Systems

A traditional way of looking at recommender system is the following: for user X recommend a set of items that maximize the user utility. As recommender systems became embedded in many online applications (e.g. online shopping) the set of recommended items has impact on the utility of many stakeholders: the content creator, the vendor, the consumer, others.

Analyze the existing work on multi-stakeholder RS. Pick an application (e.g. music recommender system) and devise a method that optimizes the recommendations by taking into account the utility for several stakeholders.

References:

- https://www.researchgate.net/profile/Himan_Abdollahpouri2/publication/338516177_Multistakeholder_recommendation_Survey_and_research_directions/links/5e191200299bf10bc3a34635/Multistakeholder-recommendation-Survey-and-research-directions.pdf

DS19 Predicting Creditworthiness

Decision support systems in banks that help the bank clerk to decide whether to give a loan to a customer or not are of help to shorten the decision time but can also fail by generating both false positives (give a loan to someone who is not able to repay it) and false negatives (deny a loan to someone who is able to repay it). Study the related work and devise a method for assessing the credit score of people.

Pick a source for your features (e.g. bank statements, social media etc.) and build a predictive model of creditworthiness

References:

- https://www.sciencedirect.com/science/article/pii/S0950705116300156

DS20 Comparing inter-generation athletes

Who is a better tennis player? Bjoern Borg or Pete Sampras? They both got 10+ grand slams but played in different eras. Who is a better footballer? Maradona or Neymar? Who is the best formula one driver? Which coaches account for better results?

It is generally considered that it is impossible to compare performers that did not compete with each other. However, it has been shown that statistical methods can be used to account for the different competing environments.

References:

- Formula for success: Multilevel modelling of Formula One Driver and Constructor performance, 1950–2014

DS21 Addressing the popularity bias in recommender Systems

In recommender system we usually want to recommend items that are personalized to the taste of the end user. However it often happens that mostly items that are popular but not necessarily of high utility get recommended. Collect a dataset and devise a method for mitigating the popularity bias.

References

- https://web-ainf.aau.at/pub/jannach/files/Conference_UMAP2016.pdf
- https://arxiv.org/pdf/2007.13019

DS22 Predicting Winners in Tennis Matches

Tennis offers dataset with details about the matches between players. Analyse such datasets and devise a feature engineering/machine learning method for predicting winners of tennis matches.

References:

- https://www.researchgate.net/publication/339103858_Predicting_tennis_match-winner_and_comparing_bookmakers_odds_using_machine_learning_techniques
- https://www.semanticscholar.org/paper/Machine-Learning-for-Professional-Tennis-Match-and-Cornman-Spellman/a8d31de35f34ed4bcd64ae62ac8392fdd87535c0?p2df
- http://cs229.stanford.edu/proj2017/final-reports/5242116.pdf

DS23 Predicting/Modeling the Popularity of Social Media Posts

Which are the characteristics of a social media post (tweet, picture, etc) that make it become popular. What is the dynamics of popularity (short/long term)? Collect a dataset and devise a model for predicting if a post is going to be popular or not.

DS24 Predicting goal scores in football

Similar to tennis match prediction.

References:

- http://cs229.stanford.edu/proj2019spr/report/46.pdf

DS25 User Identification through user interaction

Can we identify who a user is through his keyboard typing style? Mouse movements? Other unobtrusive approaches? Collect a dataset of several users and devise a feature engineering/machine learning model for identity detection.

References:

- http://cs229.stanford.edu/proj2019spr/report/48.pdf

DS26 Predicting Political Elections Outcomes From Social Media

In the period before political elections there is a lot of discussion going on on social media. Collect a dataset and devise a feature engineering/machine learning model that predicts the election outcomes from social media activity. Explain which feature values accounted for the outcomes.

DS27 Detecting author of text/Text Style analysis

Often authors of texts (emails, short stories, news articles) are anonymous. On the other hand evidence shows that each author has a distinctive style that is manifested through the language used. Apply NLP techniques to extract features that can be used in a machine learning model for the identification of the author.

DS28 Fake News Detection

News are shaping our decision making. The idea is that our decisions should be informed. However, fake news can disrupt this chain and lead to decisions that we may regret. Define what fake news are. Build a labelled dataset and a feature engineering/machine learning model for the detection of fake news. Furthermore, describe the characteristics of fake news.

Subtopic: Dis/Misinformation Mining from Social Media (from https://www.journals.elsevier.com/information-processing-and-management/call-for-papers/special-issue-on-dismisinformation-mining-from-social-media)

In the last 10 years, the dissemination and use of social media have grown significantly worldwide. Online social media have billions of users and are able to record hundreds of data from each of its users. The wide adoption of social media has resulted in an ocean of data which presents an interesting opportunity for performing data mining and knowledge discovery in a real-world context. The enormity and high variance of the information that propagates through large user communities influences the public discourse in society and sets trends and agendas in topics that range from marketing, education, business and medicine to politics, technology and the entertainment industry. This influence can however act as a double-edged sword, since it can also introduce threats to the community, if it is rooted in dissemination of disinformation, i.e. purposefully manipulated news and information, or misinformation, i.e. false and incorrect information, on social media. In recent years, the potential threats of dis/misinformation have been the subject of huge controversy in different domains like public healthcare systems, socioeconomics, business and politics. For instance, the circulation of scientifically invalid information and news can negatively affect the way the public responds to the outbreak of a pandemic disease, like COVID-19. Threats can also be posed to the legitimacy of an election system by enabling opponent campaigns to shape the public opinion based on conspiracy theories stemmed from false information. Mining the contents of social media to recognize the instances of misinformation and disinformation is a very first step towards immunizing the public society against the negative impacts they could introduce.

Traditional research on dis/misinformation mining from social media mainly focuses on descriptive methods such as fake news detection and propagation analysis, malicious bot detection, fact-checking social media content, and detecting the source of claims and rumours. The main distinguishing focus of this special issue will be the use of social media data for building diagnostic, predictive and prescriptive analysis models that can be used to understand how and why dis/misinformation is created and spread, to uncover hidden and unexpected aspects of dis/misinformation content, and to recommend insightful countermeasures to restrict the circulation of dis/misinformation and alleviate their negative effects. The ultimate goal is to immunize the social media against dis/misinformation and improving the trustworthiness of the social content and the socio-economic and business systems working based on the insights mined from social media. The main focus of the special issue is on proposing models and methods for tackling dis/misinformation in real-world scenarios.

In this special issue, we solicit manuscripts from researchers and practitioners, both from academia and industry, from different disciplines such as computer science, big data mining, machine learning, social network analysis and other related areas to share their ideas and research achievements in order to deliver technology and solutions for mining dis/misinformation from social media.

Topics of Interest

We solicit original, unpublished and innovative research work on all aspects around, but not limited to, the following themes:

- Descriptive models on fake new and malicious bot detection.
- Explainable AI for detection of dis/misinformation.
- User behaviour analysis and susceptibility prediction with regard to dis/misinformation in social media.
- Trust and reputation in social media.
- Dis/misinformation propagation modeling and trace analysis.
- Prescriptive countermeasure methods against formation and circulation of misinformation
- Predicting misinformation and bias in news on social media.
- Predictive models for early detection of hoax spread in social media.
- Social influence analysis on online social media including discovering influential users and social influence maximization.
- Assessing the influence of fake news on advertising and viral marketing in social media.
- New datasets and evaluation methodologies to help predicting dis/misinformation in social media
- User modeling and social media including predicting daily activities, recurring events Determining user similarities, trustworthiness and reliability.
- Social media and information/knowledge dissemination such as topic and trend prediction, prediction of information diffusion patterns, and identification of causality and correlation between events/topics/communities.
- Merging internal (proprietary) data with social data.

References:

- https://en.wikipedia.org/wiki/Detecting_fake_news_online

DS29: Detecting Traits from Social Media

Personality detection from social media is an established research area. The Five Factor Model has been extracted from user digital traces, such as Facebook, twitter, Instagram etc. However, there are many other traits, such as need for cognition, need for affect, need for power, the dark triad, maximization/satisfaction, eudaimonic/hedonic tendency that have not been investigated.

Conduct a user study to collect the data needed and devise a computational predictive model.

DS30: Emotions in Explanations

Co-supervision w Nava Tintarev

Explaining recommendations is an important aspect in recommender systems. Research has mostly focused on how to provide explanations to end users that are logical. However, the affective path of the user being exposed to explanations has not been investigated yet. A similar scenario is explanations given by a doctor to a patient: although the explanation might seem reasonable to an external person, the patient might be in an affective state of high arousal (shocked/surprised by learning about having a disease) and the explanation is not effective.

Devise a model of how emotions affect the perception of explanations (study related work, identify variables). Design an experiment (user study). Analyze results.

DS31: Emotions within Music Sessions

Users listen to music in sessions. Is their mood/emotional state stable or it varies within sessions? Does the user actively change music for mood regulation?

Dataset from Markus Schedl (MFM Last.FM)

http://www.cp.jku.at/people/schedl/Research/Publications/pdf/schedl_icmr_2016.pdf

DS32: Effects of events on Auction Prices

Online auction sites, such as ebay.com, offer an auction-like marketplace. In auctions, prices are driven by the demand, i.e. by how much people are willing to pay for an object. A user of these sites can observe that the prices, for which similar objects are sold vary quite a lot. The research question is if there are predictable reasons in the form of events, such as holidays season, big sports events, etc., that have an effect on the prices of sold items.

The student should find or collect an appropriate dataset, preferably over several years, with a lot of items from diverse categories. The student should perform an analysis of the data and train predictive models. Finally, the student should conclude her work with some explanations and guidelines for sellers/buyers (e.g. which time of the year is the best to sell/buy a certain kind of item).

DS33: something with football data

https://github.com/statsbomb/open-data

DS34 Streetonomics

City names are given after importat figures. Different cities have different strategies that reflect the local culture. The goal of this research is to replicate the work done on world famous cities within the context of Slovenia.

Research questions:

- are namings gender biased?
- are there street names after foreigners?
- what are the popular backgrounds (artist, politicians ...)
- how does these things change through time?

Reference
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0252869

DS35 Extraction from text

In the first step this project aims at the extraction of sarcastic/metaphorical/polemical/rhetorical/euphemistic phrases from texts. The student should identify a suitable dataset and devise an NLP/ML pipeline for the extraction of the above. IT is expected that NLP methods will be used for feature extraction (TFIDF, embeddings...) and ML techniques for training (classifiers, evluation etc.).

In the second step, the student should use the pretrained model above to monitor the activities in the social media: e.g. radicalization of users etc.