Exam in News and Market Sentiment Analytics

Exam question

This exam is of a rather open ended nature. You are tasked with the following:

> Demonstrate the skills you have learned in this course to gain relevant and useful insights into *one* of the following three sources of data **A**, **B**, or **C**.

Each of these data sources are described in detail below. You are required to base your project on one of these data sources. Please feel free to work on your own idea with the data or one of the suggested ideas described below.

Practical matters and evaluation criteria

You have one week to work on this. You must submit your answer via www.digitaleksamen.sdu.dk. The answer should be in the form of **one** pdf file containing:

- A paper describing: The **objective** of your project, the **NLP-tools** you used, and why you think they are **appropriate**, a **demonstration of the results** and **how well it performs**.
- Documentation: **A well-documented appendix with code** or (preferably) **a link to a well-documented GitHub-repository with the project's code**.

Remember to reference literature you rely on and follow good academic practice. Cite all sources, including generative AI (which you are encouraged to use if you find it relevant).

The paper itself (excluding references, appendices, code and documentation) should not exceed 15 normal pages of 2400 characters. Each table and illustration counts as 400 characters. You should not aim to fill out the maximum. You are not awarded for producing lots of text. On the contrary. Depending on your project, you might be able to convey the important details in 5 pages or less and in this case, this is what you should do.

The criteria your project is evaluated on is based on the following (in ranked order):

1. Skillful application of a few relevant NLP tools
2. Clarity of analysis, insights and the obtained results
3. Appropriateness of the ways in which performance is measured
4. Coherence between stated objectives and the chosen tools
5. Clarity of the written code and its documentation
6. Innovativeness: If you do something difficult and/or innovative then that itself can give you a few (but only a few) extra points.

Please identify yourself on the paper with your full name and SDU email address. Do not include exam number.

Working with large data

To the extend feasible you are encouraged to work with as much data as possible. But please do not hesitate to work with only a subset of the provided data if this is necessary. (And a piece of advice is to make this decision on the first day). You will not have points deducted for working with a subset of the data. But please argue for the appropriateness of how you subset/sample the data for your specific objectives.

If you encounter trouble with the data, this is probably intended to be a part of the exam.

A. S&P 500 and financial news

The following link contains data for news headlines between 2008 and 2024 paired with the closing price of the Standard & Poors 500 index:
https://www.kaggle.com/datasets/dyutidasmahaptra/s-and-p-500-with-financial-news-headlines-20082024.
You are hired by a company specializing in algorithmic trading, milli-seconds counts and high accuracy counts.
Ideas:

- Build a pipeline for extracting the financial sentiment. Can your model predict future stock market prices better than a relevant alternative? Come up with a strong evaluation strategy.
- If milliseconds counts, you might want to be able to do inference faster. Test whether a bag-of-words approach to predict sentiment as well as with e.g. a modern transformer model. To do this you might want to build your own financial sentiment dictionary.

B. Arxiv data

Data from arxiv:

- https://huggingface.co/datasets/common-pile/arxiv_abstracts_filtered
- Ideas:
  - Innovations: One definition of innovation is that an idea is close to the future but far away from the past. Construct a measure of innovation, validate its performance and show your findings
  - **Research-driven thematic sentiment index**: Pick a theme with clear market exposure (e.g. AI, quantum computing, climate tech). Use arxiv abstracts to build a time-varying “research optimism” index for that theme (e.g. via sentiment, novelty, or density of new papers) and test how e.g. NVIDIA prices relate to research optimism
  - **Diffusion of methods across fields**: Use text similarity or topic models on abstracts to track how quickly new methods (e.g. transformers, diffusion models, graph neural networks) spread from CS/ML to finance, economics or other application fields. Relate diffusion speed and breadth to investment flows or

valuation changes in sectors that rely on these methods. (**Hint:** You need to measure fields - e.g. via topic modelling or scraping the field from the arxiv website)

C. Fake and real news

Data from news media: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset The dataset consists of two CSV files, `True.csv` and `Fake.csv`, which signifies real and fake news.

You are hired by a **risk and media analytics team** in a large investment firm. They want to understand how misinformation differs from real news in tone, style and topics – and how that might distort sentiment-based trading or reputation risk models.

Ideas:

- What subjects are associated with fake news?
- Can you build a robust classifier of real/fake news? Can you find a good way of validating that it works?