Project: IDS_fakenews, https://github.com/carolinlyybek/ids_fakenews

Team: Carolin Lüübek, Katre Tiku, Liisa Tatar

Business understanding

- Identifying your business goals
  - Background
  - Business goals
  - Business success criteria
- Assessing your situation
  - Inventory of resources
  - Requirements, assumptions, and constraints
  - Risks and contingencies
  - Terminology
  - Costs and benefits
- Defining your data-mining goals
  - Data-mining goals
  - Data-mining success criteria

Business goals

Background

Our project is a part of University of Tartu's course Introduction to Data Science, where conducting a project is mandatory.
Our team chose the topic of fake news, because it's a topical and common problem in today's world. Defining which news are fake and which news are real can be hard for a person who does not indulge in news-reading, but just wants to get important information as quickly as possible.

Business goals

Our project is not meant to benefit any business, so here we state the goals from our own perspective. The goal is to embed and put to practise the skills we have gained during the course. According to the topic of the project, the goal is also to build tools for ourselves so we can be more mindful as people.

Business success criteria

It might be stated that the actual success of our project depends on the final grading of the project, but the interest and praise of our classmates and course instructors would also make the result of the project more valuable in our eyes.

Assessing the situation

Inventory of resources

The main resources of our project are the ‘data-scientists’ (our team members) with the time and effort we put in the project. The expert knowledge and insight comes from the course administrators with all the information they have already provided for us and with any answers for questions we might encounter.
In terms of software we are using Jupyter Notebook (and all the Python libraries related to data analyses) to analyse the data and reach the goals of our project.
Data is described in detail in the report’s chapter “Data understanding”, but we plan on using three datasets of news articles from Kaggle and the data we get from scraping the www.snopes.com archive.

Requirements, assumptions and constraints

The project has to be completed by December 14. On Thursday, December 17 there is a poster session held for presenting the projects in Zoom.
The datasets are public, so there will not be any legal or security obligations.
Acceptable finished work includes at least one model for predicting if the article is fake by the content (title and body text) of the article, a collection of most used words in the titles of fake news articles and of most popular categories in which these fake articles are written.

Risks and contingencies

The project completion could be delayed in case one or many members of our team encounter the problems of procrastination or any technical issues. In case of procrastination the only solution is to work as intensely as possible before the project deadline falls. With technical issues we could find help from the internet, our course mates or the course instructors.

Terminology

Fake news - false or misleading information presented as news

NLP (natural language processing) - a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data

Costs and benefits

Not relevant for our project.

Data-mining goals

Data-mining goals

We intend to create models that can predict if a news article is fake or real, which helps any person to be more mindful and intelligent. If the realness of the article is determined, people can assess further problems of why this individual wrote a fake news article and what bits of fake information go around in the world.

We plan on building models which can predict the latter on just the titles of the articles as well as the body text of the article. To make fake news detection easier for a human being, we also plan to identify the most used words in the titles of the fake articles and the categories in which most fake news articles are written about.

Data-mining success criteria

At first we measure the effectiveness of our models on our own test data, but later on we can move on to real time published articles. The success of our project depends on how well the models identify fake news articles, meaning how correctly they predict.

The success of our project also depends on how distinctive words we find that are used in fake news articles' titles and if it's possible to identify a fake news article ourselves with the knowledge we obtain.

Data understanding

- Gathering data
  - Outline data requirements
  - Verify data availability
  - Define selection criteria
- Describing data
- Exploring data
- Verifying data quality

Gathering data

Outline data requirements: Data requirements to achieve our goal is to have csv file/s of fake and real news articles that were released around the same time period and preferably published after 2016 so the articles are not as outdated. The data of the articles should contain the title, text, date created and news category/subject which are relevant for looking for patterns.

Verify data availability: 2 suitable datasets one for fake and another for real news are available from Kaggle. Since there might not be enough data and the articles are only from 2017 and it would be better to have more variety we have thought that alternative is to get extra data. One of our options is to use another dataset from Kaggle which features articles from 2016 with labels whether they were true or false. Upon researching our options we managed to find fake news research website Discourse Processing Lab, which also offers scraping fact-checking websites for research purposes. While web scraping usually is not ideal it could be used to get some extra data that we could use for testing. The scraped data also meets most of the requirements for our project and could be turned to match the format of the first two Kaggle datasets.

Define selection criteria: Our first 2 datasets are from Kaggle Fake and real news dataset. The real and fake articles are split apart so there is true.csv and fake.csv which consist of the corresponding articles. Both of those have 4 columns: title,text,subject, date, which is exactly what we need in our project.

The second Kaggle dataset is from Source based Fake News Classification dataset, which has news_articles.csv. It has more columns than necessary for our project but the relevant columns are: published, title, text, type and label.

Third dataset would come from using fake news research website [Discourse Processing Lab] ‘s fact-checking website scraping. We would use the option to scrape the entire archive of snopes.com since it seemed quite popular for getting fact ratings on articles. The csv produced from that has relevant columns like fact rating, article title, article category, article date, article origin url.

Describing Data

In our main dataset, which is the Fake and and real news dataset we have 21417 real and 23481 fake news. Real news has 20826 unique titles and 21192 unique texts while fake news has 17903 and 17455. As mentioned above those datasets have 4 columns: title,text,subject, date. Title includes the article’s title, text includes the whole text of an article, subject shows its category and date displays when the article was posted.

The second Kaggle dataset has 2096 entries for fake and real articles, where 1294 are fake and 801 are real news. Dataset has 10 columns: author, published, title, text, language, site url, main img url, type, label, title without stopwords. Where author displays the author of the article, published shows the date and time it was posted, title is the title of the article, text is the full text of the article, language displays the language of the article, site and main image url link to the corresponding url-s, type shows article type, label lists whether article is true or false and title without stopwords is title without stop words which means that words which are considered stop words like “a”, “that”, “so” etc are filtered out before processing of natural language data. For our project the only relevant columns are: published, title, text, type and label.

Third dataset which originates from snopes.com fact-checking website has 37957 entries of data. It has 9 columns : fact rating phase1, snopes url, article title, article category, article date, article claim, article origin url, index paragraph, page is first citation. Fact rating list whether article is true, false, mixture etc, snopes url link to the snopes website where the article was rated, article title shows articles title, category shows its category, date shows when it was posted, article claim shows what the article claimed, origin url link to the original website of the article, index paragraph shows the index by paragraph, page is first citation tells whether the article is first citation or not. Relevant columns for us are: fact rating, article title, article category, article date, article origin url.

The first two Kaggle sets will be easier to merge since they have quite similar columns, the first Kaggle dataset needs just a label column for indicating whether an article is true or false and from the second dataset we could drop the unnecessary columns. Third dataset will be harder to merge since it’s fact rating has more values than just true or false and doesn’t directly include the whole article’s text. Some additional data preparation will be needed to do to match it to our Kaggle datasets.

Exploring data

As mentioned before our main dataset has 21417 real and 23481 fake news. Real news has 20826 unique titles and 21192 unique texts while fake news has 17903 and 17455. The image below shows the difference in unique values between real and fake news.

```
In [21]: real_data.nunique()

Out[21]: title      20826
         text       21192
         subject        2
         date         716
         dtype: int64
```

```
In [20]: fake_data.nunique()

Out[20]: title      17903
         text       17455
         subject        6
         date        1681
         dtype: int64
```

Both of those datasets have no missing values as we can see from the image below.

```
In [6]: fake_data.count()

Out[6]: title      23481
        text       23481
        subject    23481
        date       23481
        dtype: int64
```

```
In [4]: real_data.count()

Out[4]: title      21417
        text       21417
        subject    21417
        date       21417
        dtype: int64
```

The second Kaggle set has 2096 entries for fake and real articles, where 1294 are fake and 801 are real news. Image below shows the number of unique values for each column.

```
In [22]: articles.nunique()

Out[22]: author                      491
         published                  2006
         title                      1784
         text                       1941
         language                      5
         site_url                     68
         main_img_url               1229
         type                          8
         label                         2
         title_without_stopwords    1780
         text_without_stopwords     1937
         hasImage                      2
         dtype: int64
```

We can also notice that some values are missing from the columns like 46 values from text and 49 values from text without stop words if we look at the image below.

```
In [10]: articles.count()

Out[10]: author                      2096
         published                   2096
         title                       2096
         text                        2050
         language                    2095
         site_url                    2095
         main_img_url                2095
         type                        2095
         label                       2095
         title_without_stopwords     2094
         text_without_stopwords      2046
         hasImage                    2095
         dtype: int64
```

The third dataset has 37957 entries of data and the image below shows the unique values for each column.

```
In [23]: snopes.nunique()

Out[23]: fact_rating_phase1                 12
         snopes_url_phase1                5640
         article_title_phase1             5638
         article_category_phase1           324
         article_date_phase1              1644
         article_claim_phase1             5638
         article_origin_url_phase1       34031
         index_paragraph_phase1            102
         page_is_first_citation_phase1       2
         dtype: int64
```

7 values from the article category column and 1 value from the origin url column are missing as we can see from the image below.

```
In [8]: snopes.count()

Out[8]: fact_rating_phase1                 37957
        snopes_url_phase1                  37957
        article_title_phase1               37957
        article_category_phase1            37950
        article_date_phase1                37957
        article_claim_phase1               37957
        article_origin_url_phase1          37956
        index_paragraph_phase1             37957
        page_is_first_citation_phase1      37957
        dtype: int64
```

Verifying data quality

The most concerning factor with all of our datasets is that most of them have less unique titles and texts than we have actual lines of data which mean that some of the data is repeating. To tackle that issue is to remove the duplicates. The first Kaggle dataset doesn’t have a problem with missing values while the second Kaggle dataset and snopes dataset do. Since it’s less than 100 values in both of those cases it’s not that big of an issue and for some of them we could look into the original articles or link that were provided to find the missing data and if that’s not possible another solution is to remove those entries.

Planning the project

Detailed plan

1. **Preparation**
   At this stage, we focus on understanding project goals. We also collect relevant and accurate data from our data sources. At this stage we also take the project to pieces and analyse what needs to be done and how much time will we spend on doing them.
   This stage is time consuming and will approximately take up about 25% of all time we spend doing the project.

2. **Understanding the Data**
   This step is about making sure what data we collected and if it is relevant to our project.At this step we also transform and prepare data for the next step.
   This step should take up 15% of time.

3. **Data preparation**
   The goal of the stage is to prepare data and assess its suitability. This step contains identifying data quality problems, discovering the insight from data to formulate hypotheses regarding hidden information. At this step it is important to conduct a data mining exercise in which we can verify assumptions and see if we understand the data correctly.
   This step will take up approximately 30% of time.

4. **Modeling**
   Statistical models are built, selected and checked during this stage. Since some techniques have specific data requirements, we might have to go back to the data preparation stage. In this stage we should conduct a few experiments with a small amount of data before dealing with the whole data.
   This step will probably take up 20% of time.

5. **Performance evaluation**
   This step evaluates what is performing well and poor, in order to check model assumptions and identify improvement. This means testing all the models that we have done.
   It will approximately take up 10% of time.

List of methods and tools

We are using Python as the main language for data processing. We will mainly do all of our work in Jupyter Notebook and use all the Python libraries related to data analyses there. We are using three datasets of news articles from Kaggle, in addition to that we are also using the data we get from www.snopes.com archive. We will be using NLP(natural language processing) and classification methods to train a detection model.