

## Task description

# Break the FAKE

## HackYeah programmer!

Imagine, every day you are getting an email with a newspaper, but not the same as everyone. This newspaper was written specifically for you and contains real, verified and current economical information from multiple reputable sources. Do you want to leave a mark on Polish reality?

Help us with providing credible information and stop fake news. Take the challenge and Break the Fake

### About our task

The growing amount of information generates more and more noise, making it difficult to catch fake news. In this edition, we expect you to design a solution that can structure daily news, as well as evaluate its sentiment and determine the probability that a piece of information is false. This task will test your practical programming skills, and a lot of freedom in its implementation will not limit your creativity.

We expect you to design a solution, based on Python and/or R languages, which any user will be able to operate. It has to allow browsing articles written in Polish and coming from multiple sources. The solution must classify each article based on its content into at least one of the categories presented in the appendix (1).

The conditions your solution must meet are specified in two categories - necessary and bonus.

### Our requirements

I      The application should (necessary conditions):

1. have as simple, friendly, functional and intuitive interface as possible, allowing filtering by date, category, source, sentiment rating, likelihood of fake news.
2. be useable on Windows 10 operating system
3. display article items containing following information:
  - a) title,
  - b) abstract,
  - c) source (source name and link),

- d) time of publication,
  - e) author of the article,
  - f) tags,
  - g) category (classified by algorithm),
  - h) sentiment assessment (positive, neutral, negative),
  - i) fake news probability assessment (up to 5 categories).
4. archive the collected articles.
  5. allow to set the frequency of updating the archive of articles.

The provided solution must be based on tools that will not incur additional costs for further use by the Ministry of Finance.

**II** The application may provide additional functions that significantly affect the functionality of the interface (bonus conditions), e.g.

- the ability to create charts for individual filtering possibilities,
- determine the original source of the article, including graphical visualization,
- advanced sentiment assessment, e.g. emotions,
- probability of hate speech.

### **Evaluatoin of your projects**

We will evaluate the prepared interface, which will be presented before the committee by the Contest participant or Team. The solution presented should use articles from a maximum of the last 30 days. The evaluation consists of two stages.

#### **Stage 1**

We check if the prerequisites are met, which determines qualification for Stage II of the evaluation.

#### **Stage 2**

We evaluate the quality of the delivered solution with additional functionalities on a scale from 0 to 100 points, consisting of:

- a) the degree of implementation and quality of the individual component functionalities of the solution (0-40). In particular, the evaluation criteria will take into account the following elements: the number of data sources, the number of articles, the number and accuracy of article classification, the number and accuracy of sentiment assessment, the number and accuracy of fake news probability assessment,

- b) intuitiveness of the designed interface and its ergonomics (0-20). The evaluation criteria will take into account the user-friendliness for a user with basic computer skills and the aesthetics of the interface,
- c) architecture of the solution and the ability to implement the system in a production environment: scoring (0 to 20). The scoring will take into account the evaluation of such elements as the composition of the system modules, scalability, flexibility with respect to different data sources, issues of monitoring the work of the application, performance, accountability, code portability,
- d) the level of complexity of the sentiment classification method used and the evaluation of the article as fake news (0-10). The evaluation criteria will take into account the adequacy of the classification method used and its sophistication,
- e) provision of additional features not indicated in the requirements that significantly affect the functionality of the interface (0-10), such as the ability to create summary charts for each category of filters, determination of the original origin of the article (including graphical visualization), determination of sentiment in an advanced way, such as emotions.

### Additional information

All solution files must be provided as a .zip file named "Break-the-FAKE-*teamnumber*" and contain at least the following:

- a) a file named "INFO.zip" - archive containing all the information about the provided solution and the possible configuration of the environment, as well as a description of the steps to run the interface,
- b) a file named "CODE.zip" - archive containing source code or source code file locations allowing to obtain a compiled, working version of the software,
- c) a file named "DASHBOARD.zip" - archive containing a compiled, version of the tool, based on the source codes contained in the file "CODE.ZIP" or a link with its location,
- d) a file named "OTHERS.zip" - archive containing any additional information.

## Break the FAKE

### Programisto eventu HackYeah!

Wyobraź sobie, że codziennie na skrzynkę e-mail dostajesz gazetę, ale nie taką samą jak wszyscy. Ta została napisana specjalnie dla Ciebie i zawiera tylko prawdziwe, zweryfikowane i aktualne informacje gospodarcze i ekonomiczne z wielu źródeł, którym możesz zaufać. Chcesz odcisnąć ślad na polskiej rzeczywistości? Pomóż nam dostarczać wiarygodne informacje i zatrzymaj fake newsy. Podejmij wyzwanie i **Break the Fake!**

### O naszym zadaniu

Rosnąca ilość informacji generuje coraz to większy szum utrudniający wyłapywanie fake newsów. W tej edycji oczekujemy od Was zaprojektowania rozwiązania, które będzie w stanie ustrukturyzować codzienne newsy, a także ocenić ich sentyment oraz określić prawdopodobieństwo, że dana informacja jest nieprawdziwa. To zadanie sprawdzi praktyczne umiejętności programowania, a duża swoboda w jego realizacji nie będzie ograniczać Waszej kreatywności.

Oczekujemy zaprojektowania rozwiązania, bazującego na językach Python i/lub R, które będzie potrafił obsługiwać każdy użytkownik. Ma ono umożliwić przeglądanie artykułów napisanych w języku polskim i pochodzących z wielu źródeł. **Rozwiązanie musi klasyfikować każdy artykuł na podstawie jego treści do co najmniej jednej kategorii przedstawionych załączniku (1).**

Warunki jakie musi spełnić projektowane przez Was rozwiązanie zostały określone w dwóch kategoriach – konieczne i dodatkowe.

### Nasze wymagania

#### I Aplikacja powinna (warunki konieczne):

6. Posiadać możliwie jak najprostszy, przyjazny, funkcjonalny i jak najbardziej intuicyjny interfejs, umożliwiający filtrowanie według dat, kategorii, źródła, oceny sentymentu, prawdopodobieństwa *fake newsa*.
7. Być dostosowana do zastosowania na systemie operacyjnym Windows 10.
8. Wyświetlać pozycje artykułów wraz z następującymi informacjami:

- j) tytuł,
- k) streszczenie,

- l) źródło artykułu (nazwa źródła oraz link),
  - m) czas publikacji,
  - n) autora artykułu,
  - o) tagi,
  - p) kategorię (dokonaną przez algorytm klasyfikacyjny),
  - q) określenie sentymentu dotyczący przekazu (pozytywny, neutralny, negatywny),
  - r) ocenę prawdopodobieństwa, że artykuł jest *fake newsem* (maksymalnie do 5 kategorii).
9. archiwizować zebrane artykuły.
10. umożliwić ustawienie częstości zasilania archiwum artykułów.

Dostarczone rozwiązanie musi bazować na narzędziach, których dalsze wykorzystanie przez Ministerstwo Finansów nie będzie wiązało się z ponoszeniem dodatkowych kosztów.

**II Aplikacja może zapewniać dodatkowe funkcje, które w istotny sposób wpływają na funkcjonalność interfejsu (warunki dodatkowe) np.**

- możliwość tworzenia wykresów dla poszczególnych możliwości filtrowania,
- określenie pierwotnego pochodzenia artykułu, w tym grafowa wizualizacja,
- określenie sentymentu w sposób zaawansowany np. emocje,
- ocena prawdopodobieństwa, że w artykule znajduje się „mowa nienawiści”.

### **Ocena Waszych projektów**

Oceniany jest przygotowany interfejs, który uczestnik Konkursu lub Zespół zaprezentuje przed komisją. Do oceny prezentowanego rozwiązania interfejs ma przedstawiać artykuły z maksimum ostatnich 30 dni. Ocena składa się z dwóch etapów.

#### **Etap 1**

Oceniane jest spełnienie warunków koniecznych, a następnie podejmowana jest decyzja o zakwalifikowaniu rozwiązania do etapu II oceny.

#### **Etap 2**

Oceniana jest jakość dostarczonego rozwiązania wraz z dodatkowymi funkcjonalnościami na skali punktowej od 0 do 100 punktów, na które składają się:

- f) stopień realizacji i jakość poszczególnych składowych funkcjonalności rozwiązania (0 - 40). Kryteria oceny będą uwzględniały w szczególności elementy: liczba źródeł danych, liczba artykułów, liczba i trafność klasyfikacji artykułu, liczba i trafność określenia sentymentu artykułów, liczba i trafność oceny prawdopodobieństwa artykułu jako *fake newsa*,

- g) intuicyjność obsługi zaprojektowanego interfejsu i jego ergonomia (0 - 20). Kryteria oceny będą uwzględniały przyjazność obsługi dla użytkownika o podstawowych umiejętnościach korzystania z komputera oraz estetykę interfejsu,
- h) architektura rozwiązania oraz możliwość wdrożenia systemu w środowisku produkcyjnym: punktacja (0 - 20). Punktacja będzie uwzględniała ocenę takich elementów jak: kompozycja modułów systemu, skalowalność, elastyczność względem różnych źródeł danych, zagadnienia dotyczące monitoringu pracy aplikacji, wydajności, rozliczalności, przenośności kodu,
- i) poziom złożoności wykorzystanej metody klasyfikacji sentymentu oraz oceny artykułu jako *fake newsa* (0 - 10). Kryteria oceny będą uwzględniały adekwatność zastosowanej metody klasyfikacji oraz jej zaawansowania,
- j) zapewnienie dodatkowych funkcji niewskazanych w wymaganiach, które w istotny sposób wpływają na funkcjonalność interfejsu (0 - 10), np. możliwość tworzenia wykresów podsumowujących dla poszczególnych kategorii filtrów, określenie pierwotnego pochodzenia artykułu (w tym grafowa wizualizacja), określenie sentymentu w sposób zaawansowany np. emocje.

### Informacje dodatkowe

Wszystkie pliki dotyczące rozwiązania muszą zostać dostarczone w formie pliku .zip o nazwie „Break-the-FAKE-numerzespołu” i zawierać co najmniej poniższe elementy:

- e) plik o nazwie „INFO.zip” – archiwum zawierające wszelkie informacje dotyczące dostarczonego rozwiązania oraz ewentualnej konfiguracji środowiska, a także opis kroków umożliwiających uruchomienie interfejsu,
- f) plik o nazwie „CODE.zip” – archiwum zawierające kody źródłowe lub adresy lokalizacji plików z kodami źródłowymi pozwalające na uzyskanie skompilowanej, działającej wersji oprogramowania,
- g) plik o nazwie „DASHBOARD.zip” – archiwum zawierające skompilowaną, na podstawie kodów źródłowych zawartych w pliku „CODE.ZIP”, wersję narzędzia lub odnośnik z jego lokalizacją,
- h) plik o nazwie „OTHERS.zip” – archiwum zawierające wszelkie inne informacje.