**PROIECT**

**Echipa D**

**Titlu**: Sentiment Analysis for Financial News/ Analiza sentimentelor pentru stirile financiare

**Colectarea corpusului :**

* <https://www.kaggle.com/ankurzing/sentiment-analysis-for-financial-news>
* <https://www.kaggle.com/shwethaspdevam/comparison-of-models-for-financial-news>
* <https://www.researchgate.net/profile/Rob_Schumaker/publication/255604315_Sentiment_Analysis_of_Financial_News_Articles/links/56d49bcb08ae2cd682b93c7c.pdf> (articol)
* <https://arxiv.org/pdf/1811.11008.pdf> (articol)
* <https://www.researchgate.net/publication/338898380_Financial_News_Sentiment_Analysis_Using_Lexicon-_Based_Labelling_and_Machine_Learning-Based_Algorithm> (articol)
* <http://ceur-ws.org/Vol-862/FEOSWp4.pdf> (articol)

**Descrierea bazei de date**

* Acest set de date conține sentimentele (parerea) pentru titlurile știrilor financiare din perspectiva unui investitor cu amanuntul ( retail investor)
* Setul de date conține două coloane: ‘Sentiment’ și ‘ News Headline’ ( Titlul știrilor).

Variabila sentiment poate conține următoarele valori: Negativ/Neutru/Pozitiv

**Dezvoltarea unui instrument de prezicere**

* Clasificatorul Naive Bayes
* Arbori de decizie
* RandomForestClassifier

**Obiectiv:** Parerea unui investitor cu amănuntul cu privire la titlurile știrilor financiare

**Surse**

* <https://www.kaggle.com/shwethaspdevam/comparison-of-models-for-financial-news>
* <https://www.kaggle.com/ankurzing/sentiment-analysis-for-financial-news>

**Arhitectura proiectului**

1. Prezentarea proiectului **(comun) –săptămâna 6**
2. Prezentarea bazei de date **(comun) –săptămâna 6**

* importarea bazei de date
* prezentarea variabilelor
* activarea pachetelor utilizate pentru analiză

1. Curățarea și prelucrarea bazei de date **- săptămâna 7**

* ștergerea virgulelor și ștergerea cuvintelor de oprire **(Maria)**
* stemming and lemmatization of text data **(Loredana)**
* eliminarea duplicatelor **(Laura)**
* codificarea variabilei categoriale ( realizam doua coloane, respectiv una cu categoria, una cu codul categoriei ) **(Madalina)**
* previzualizarea bazei de date curățate și preprocesate **(comun)**

1. Realizăm analiza exploratorie a variabilei de interes **–săptămâna 8**

* vizualizarea categoriilor variabilei de interes (prin intermediul unui grafic de tip bară) **(Maria)**

1. Realizarea unor modele de clasificare –**săptămâna 9**

* regresie logistică **(Loredana)**
* clasificatorul Gradient Boosting **(Laura)**
* Decision Tree **(Madalina)**
* clasificatorul K-nearest neighbour **(Maria)**

1. Alegerea celui mai bun model **–săptămâna 10**

* vom alege în funcție de mai multe de mai multe criterii ( de exemplu: precizia, acuratețea, stabilitatea modelului, senzitivitatea etc.) **(comun)**
* concluzii. **(comun)**