**Universitatea “Alexandru Ioan Cuza” din Iaşi**

**Facultatea de Economie şi Administrarea Afacerilor**

**Specializarea: Data Mining**

**PROIECT**

**Analiza de clasificare a sentimentelor în privința știrilor financiare**

**PROFESOR COORDONATOR: STUDENTI:**

**Prof. Univ.dr.Mihaela Plămadă Bâza Laura-Ștefania**

**Gîncianu Mădălina**

**Horaicu Maria Mădălina**

**Palade Loredana-Liliana**

**Iași, 2021**

CUPRINS

[1. Introducere 3](#_Toc62058732)

[2. Descrierea proiectului 3](#_Toc62058733)

[3. Motivația realizării proiectului – utilitatea acestuia 3](#_Toc62058734)

[4. Literatura de specialitate – Documentare 4](#_Toc62058735)

[5. Impedimente 6](#_Toc62058736)

[6.Arhitectura proiectului 6](#_Toc62058737)

[7. Datele utilizate 6](#_Toc62058738)

[8. Procesul de pre-procesare a datelor utilizate 7](#_Toc62058739)

[9. Analiza de clasificare 9](#_Toc62058740)

[9.1. Clasificarea automată 10](#_Toc62058741)

[9.2. Clasificarea manuală 11](#_Toc62058742)

[10. Clasificarea unor noi observații 12](#_Toc62058743)

[11. Contribuție proiect 12](#_Toc62058745)

[Concluzii 13](#_Toc62058746)

[Bibliografie 13](#_Toc62058747)

# 1. Introducere

Articolele de știri sau mai precis **știrile financiare** au fost întotdeauna o forță importantă în formarea imaginii financiare a unei companii în mintea publicului larg, în special a investitorilor și a traderilor. Având în vedere cantitatea mare de știri generate în ziua de azi, este posibil să se extragă sentimentul general în privința unei anumite companii, prezentat de agențiile media prin știri financiare într-un anumit timp, care poate fi utilizat pentru a evalua impactul pe termen lung asupra potențialului de investiții al companiei. **Analiza sentimentelor** este procesul de extragere a conținutului emoțional din astfel de texte iar **clasificarea sentimentelor** este o componentă importantă a analizei sentimentelor și își propune să clasifice o opinie din text ca exprimând un sentiment pozitiv, negativ sau neutru.

# 2. Descrierea proiectului

**Traderii și comercianții** caută informații despre companiile cu care vor să colaboreze, în care să investească sau de la care vor să cumpere acțiuni pentru tranzacționare pe termen lung și scurt iar în acest sens o sursă frecventă de informații sunt **știrile financiare** care oferă actualizări despre activitățile unei companii cum ar fi despre expansiunea ei, venituri mai bune / mai slabe decât se aștepta, noi produse și multe altele. În funcție de aceste știri financiare, comercianții pot determina un trend, o anumită tendință și pe baza acesteia decid dacă să investească sau nu în ea. Astfel că **identificarea sentimentelor știrilor financiare** îi ajută foarte mult pe traderi în luarea decizilor de a investi într-o companie, de a-și realiza strategii de trading dar mai ales de a lua decizii într-un timp cât mai scurt. În cadrul acestui proiect, baza de date a știrilor financiare conține și sentimentele deja identificate a acestora iar **obiectivul urmărit** este de a antrena o serie de clasificatori pe aceste date dintre care il alegem pe cel mai bun pe baza nivelului de acurateșțe pe care îl furnizează și încercăm cu ajutorul lui să **clasificăm o nouă știre financiară**.

# 3. Motivația realizării proiectului – utilitatea acestuia

Motivația din spatele acestei analize de clasificare a sentimentelor știrilor financiare este că oamenii iau decizii folosind majoritatea informațiilor disponibile. Durează de obicei câteva minute pentru a descoperi informații noi și a lua o decizie. Un algoritm este capabil să proceseze gigaocteți de text din fluxuri multi-sursă în câteva secunde. Am putea exploata această diferență pentru a crea o **strategie de trading** într-un timp cât mai scurt.

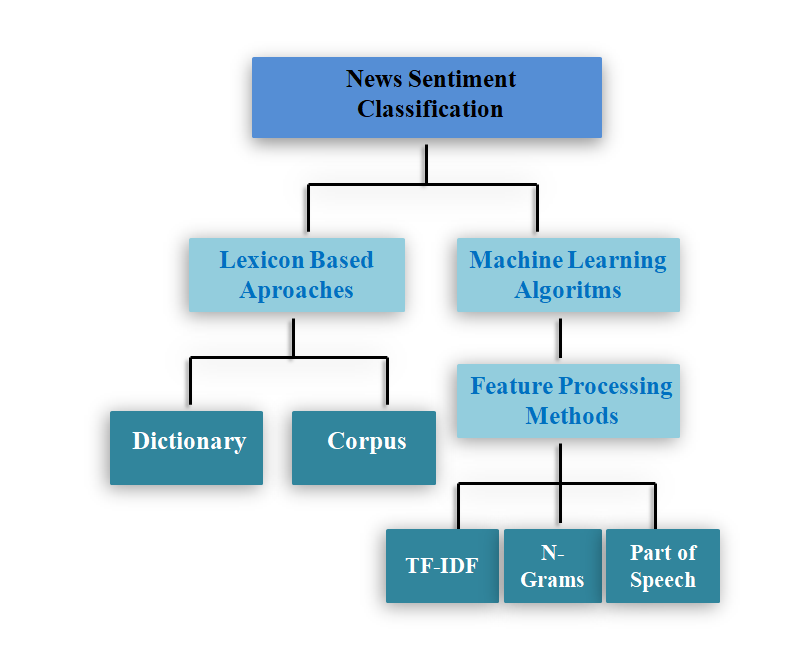
# 4. Literatura de specialitate – Documentare

Pe baza literaturii de specialitate analizate am putut observa faptul că metodele de clasificare în privința sentimentului știrilor financiare se împart în două categorii și anume:

– **metode de clasificare care se bazează pe lexicon**

– **metode care se bazează pe algoritmi de clasificare machine learning**

Astfel că în continuare am construit o schiță pe baza literaturii de specialitate a metodelor ce pot utilizate în analiza sentiment de clasificare a știrilor financiare. În cadrul acestui proiect din metodele menționate în schița realizată vor fi utilizați algoritmii de clasificare machine learning care se bazează pe una dintre cele 3 metode menționate, TF-IDF, N-Grams și Part of Speech și anume metoda cea mai cunoscută de procesare a trăsăturilor **Term Frequency-Inverse Document Frequency**(TF-IDF).

****

**Figura 1** – Metode ce pot fi utilizate în analiza sentiment

de clasificare a știrilor financiare

**Tabel 1** – *Sinteză a literaturii de specialitate privind analiza sentiment de clasificare a știrilor financiare*

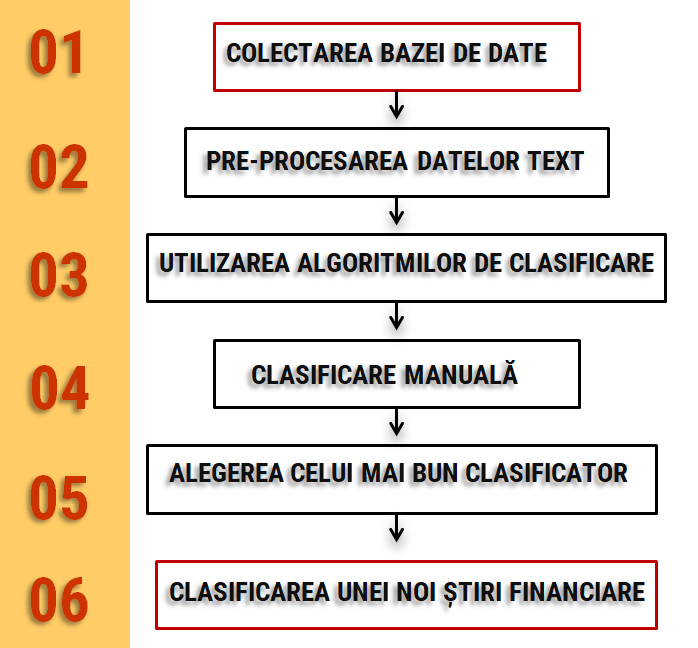
|  |  |  |  |
| --- | --- | --- | --- |
| **STUDIU** | **OBIECTIV URMĂRIT** | **ANALIZE**  **REALIZATE** | **CONSTATĂRI**  **RELEVANTE** |
| **Financial News Sentiment Analysis Using LexiconBased Labelling and Machine Learning-Based Algorithm**  *Ling Wu, Siew Hock Ow* | – a analiza sentimentul de știri financiare care pot fi utilizate încontinuare pentru a analiza impactul acestor știri financiare sentimentele cu privire la mișcarea prețului bursier. Se propune utilizarea etichetării bazată pe lexic și pe algoritmi bazați pe machine learning. | Sunt utilizați algoritmi de învățare automată supravegheată precum:  – Support Vector  – Random Forest  – Bernoulli Naive Bayes  – Complement Naive Bayes  – Multinomial Naive Bayes  – Gaussian Naive Bayes  – Stocastic Gradient Descent | – Rezultatele aceastei cercetări arată că dintre cei șapte clasificatori bazați pe algoritmi de învățare automată supravegheați folosiți și testați separat doar **clasificatorul Stocastic Gradient Descent** a atins un procent de **70% de precizie**. |
| **Sentiment Analysis of Financial News Using Performance Indicators**  *Srikumar Krishnamoorthy* | – a examina utilizarea indicatorilor de performanta pentru a prezice polaritatea in textele financiare | **Clasificator de sentimente ierarhice (HSC)** ce cuprinde**:**  – Lexicon specific domeniului- utilizeaza un dictionar LM  – Etichetarea textului utilizand lexicul Textului financiar  – Modelul clasificatorului de polaritate | – **Eficacitatea clasificatorului** a fost demonstrată printr-o analiză experimentală riguroasă pe un set de date financiare de referință Abordarea propusa s-a dovedit a fi destul de eficienta, obtinand o precizie de peste 80% si o masura F de peste 68% in toate seturile de date studiate. |
| **Semantic Based Sentiment Analysis in Financial News**  *Juana María Ruiz-Martínez, Rafael Valencia-García1, Francisco García-Sánchez* | – a realiza o analiza sentiment semantica asupra stirilor financiare, în cadrul articolului e prezentat un algoritm semantic pentru extragerea opiniilor aplicat domeniului financiar | Sunt utilizate **metode de procesare a limbajului natural** pentru a comenta stiri financiare in conformitate cu ontologia financiara, iar apoi stirile financiare comentate sunt analizate prin trecerea acestora printr-un numar de **liste de gazete** ce rezulta in doua seturi separate, un set cu stiri financiare pozitive si un set cu stiri financiare negative. | – **Rezultatele obtinute** ale analizei sentiment sunt foarte promitatoare, cu o medie a acuratetei agregate de 87%. Important este de mentionat faptul ca in privinta rezultatelor s-a luat in considerare decizia finala a sistemului si nu procesul pe care sistemul il efectueaza pentru a produce o asemenea decizie. |
| **Sentiment Analysis of Financial News Articles**  *Robert P. Schumaker , Yulei Zhang and Chun-Neng Huang* | – asocierea unui sistem de predicție a articolelor de știri financiare, **AzFinText**, cu tehnici de analiza a sentimentelor | Model Building a **AZFinText** – partiționează datele pentru a răspunde cel mai bine la întrebările de cercetare. În cazul obiectivității/subiectivității se folosește OpinionFinder. Pentru algorimtul de învățare automată s-a implementat **SVR** (Sequential Minimal Function) - functia de optimizare. Această funcție permite predicție numerică discretă în loc de clasificare. S-a selectat un nucleu liniar și de zece ori validare încrucișată | – AZFinText a fost cel mai în măsură să prezică articole subiective în Precizie Direcțională (59,0% până la 50,4%) și rentabilități de tranzacționare (3,30% - 2,41%), dar nu și apropiere (0,103 față de 0,0516). S-a simțit că subiectivitatea articolelor poate fi influențată comportamentul de tranzacționare  – AZFinText a fost cel mai în măsură să prezică articole subiective negative în Precizie Direcțională (50,9% la 50,4%) și randamente de tranzacționare (3,04% la 2,41%), dar nu și apropiere (0,0576 versus 0,0516). Se consideră că aceste rezultate pot fi atribuite investitorilor care reacționează mai puternic la articole negative.  – AZFinText a fost mai capabil să prezică scăderi în articolele pozitive (53,5%) și creșteri în articolele negative și neutre (52,4%, respectiv 49,5%). |

# 5. Impedimente

În cadrul proiectului nu am dat peste anumite provocări ci mai bine spus am avut parte de niște impedimente cum ar fi păstrarea știrilor financiare pe liniile coloanei corespunzătoare acestora în baza de date în urma realizării procesului de lemmatizare deoarece codul utilizat în programul **Python** ne furniza un rezultat de tip listă.

# 6.Arhitectura proiectului

În continuare am construit o schemă a arhitecturii proiectului care reprezintă pașii care au fost urmați și respectați în realizarea analizei sentiment de clasificare a știrilor financiare.

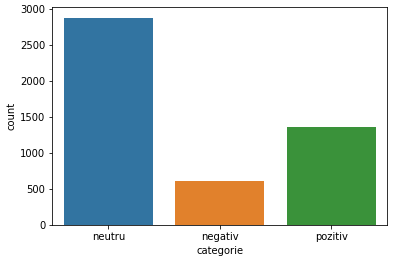
****

**Figura 2** – Arhitectura proiectului

# 7. Datele utilizate

Baza de date ***FinancialPhraseBank*** utilizată în cadrul proiectului a fost preluată de pe site-ul ***Kaggle*** și conținesentimentele pentru titlurile știrilor financiare din perspectiva unui investitor cu amănuntul. Setul de date conține 2 coloane, prima coloană, categoria, conține sentimentele deja identificate pentru titlurile de știri financiare și anume dacă sunt pozitive, negative sau neutre iar cea de a doua coloană conține titlurile acestor știri financiare.

|  |  |
| --- | --- |
| **categorie** | **număr știri** |
| negativ | 604 |
| neutru | 2879 |
| pozitiv | 1363 |



**Figura 3** – Reprezentarea grafică a știrilor pe sentimente

# 8. Etapa de pre-procesare a datelor utilizate

Etapa de pre-procesare în cadrul analizei sentiment a știrilor financiare reprezintă **procesul de reducere al zgomotului** și de preparare a textului pentru a realiza **clasificarea sentimentelor**. De asemenea, multe cuvinte în text nu au impact asupra sentimentului iar păstrarea lor face ca dimensionalitatea textului să fie o problemă și prin urmare clasificarea sentimentelor știrilor financiare mai dificilă.

Obiectivul etapei de pre-procesare a datelor este de a reduce zgomotul, de a ajuta la îmbunătățirea performanței clasificatorului și de a grăbi procesul de clasificare. Primul lucru a fost să realizăm traducerea bazei de date din română în engleză care apoi a necesitat o corectare manuală a unor greșeli de traducere. **Erori de traducere** au mai fost identificate și cu ajutorul afișării cuvintelor frecvente cum ar fi în privința cuvântului euro am identificat erori de traducere precum *eur* și, respectiv *euroo* ce au fost ulterior înlocuite cu forma corectă. De asemenea am mai identificat și cuvinte ce nu au fost identificate în urma eliminării cuvintelor de oprire astfel că am apelat la realizarea unei liste manuale cu astfel de cuvinte pe care le-am ulterior le-am eliminat. În continuare este prezentată o schiță ce reprezintă pașii care au fost urmați pentru a realiza acest proces de pre-procesare a datelor utilizate.

– procesul de împărțire în cuvinte a setului de date

**TOKENIZATION**

– eliminarea semnelor de punctuație

– înlocuirea cratimei cu spațiu

**PUNCTUATION**

**RE**

– am apelat la eliminarea numerelor din setul de date deoarece acestea nu furnizează un înteles semantic

**NUMBERS**

– transformarea literelor mari ale setului de date în litere mici

**LOWER TEXT**

– eliminarea cuvintelor de oprire din text folosind librăria **spacy**

**STOPWORDS**

– prin afișarea celor mai frecvente cuvinte conținute în setul de date s-a identificat erori de traducere și cuvinte ce nu s-au eliminat în etapa anterioară

**FREQUENT**

**WORDS**

– listă ce conține cuvinte identificate cu frequent words ce nu au fost eliminate cu stopwords

**CUSTOMISED**

**STOPWORDS**

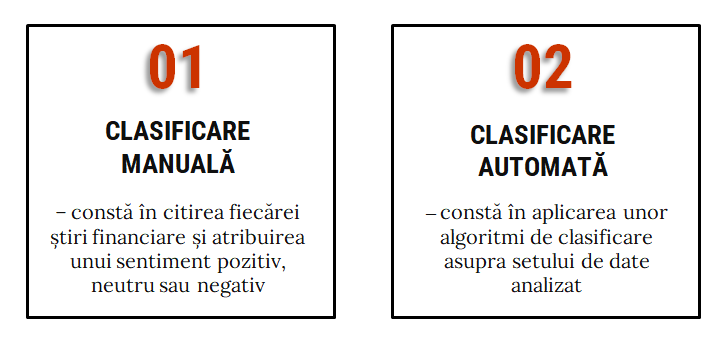
– procesul de lemmatizare realizat cu librăria spacy pe baza modelului pentru română **ro\_core\_news\_lg**

**LEMMATIZATION**

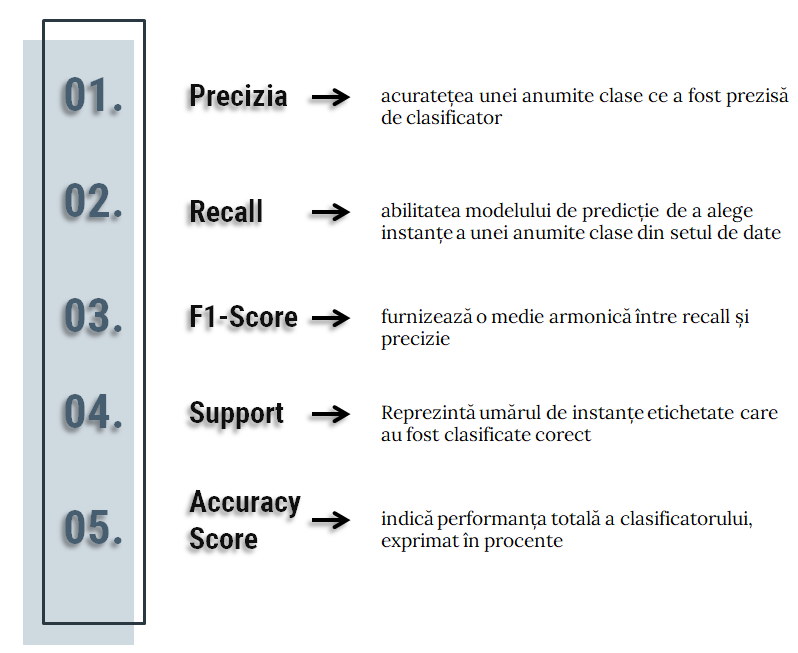
**Figura 4** – Proposed pre-processing approach process

# 9. Analiza de clasificare

Pentru a realiza clasificarea sentiment a știrilor financiare 4 metode au fost testate pentru determinarea sentimentelor: **metoda manuală** care a constat în citirea fiecărei știri financiare și atribuirea unui sentiment pozitiv, neutru sau negativ, această metodă fiind realizată doar pentru 20% din observațiile bazei de date deoarece este o metodă ce necesită mult timp la dispoziție și **metoda automată** care presupune utilizarea unor algoritmi de clasificare asupra setului de date analizat fiind utilizate în acest sens **Regresia Logistică**, **Arbori de Decizie**, **K-Means Neighbour** și **Multinomial Naive Bayes**.



**Figura 5** – Clasificare automată și clasificare manuală

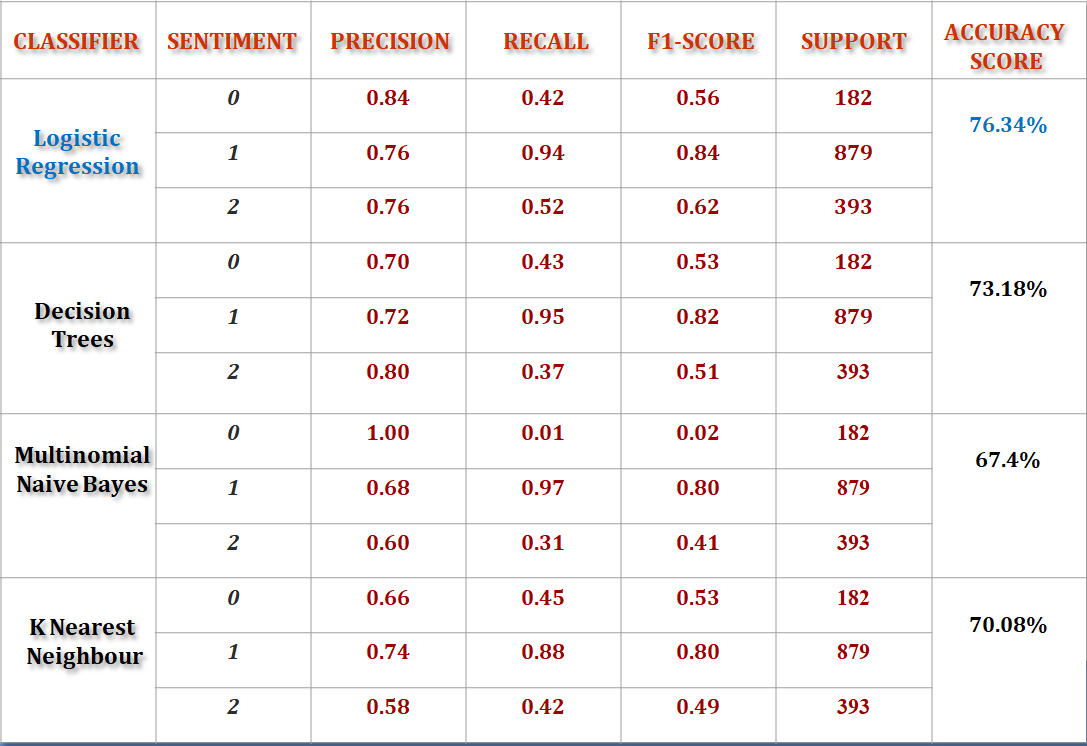


**Figura 6** – Prezentarea succintă a semnificației rezultatelor în privința clasificatorilor

# 9.1. Clasificarea automată

Tabelul realizat conține rezultate în privința a 4 clasificatori utilizați pentru știrile financiare ce aveau deja identificate sentimentele, iar pe baza comparării clasificatorii pe baza acurateței observăm că cel mai bun clasificator este reprezentat de regresia logistică ce are o acuratețe în valoare de 76.34%.

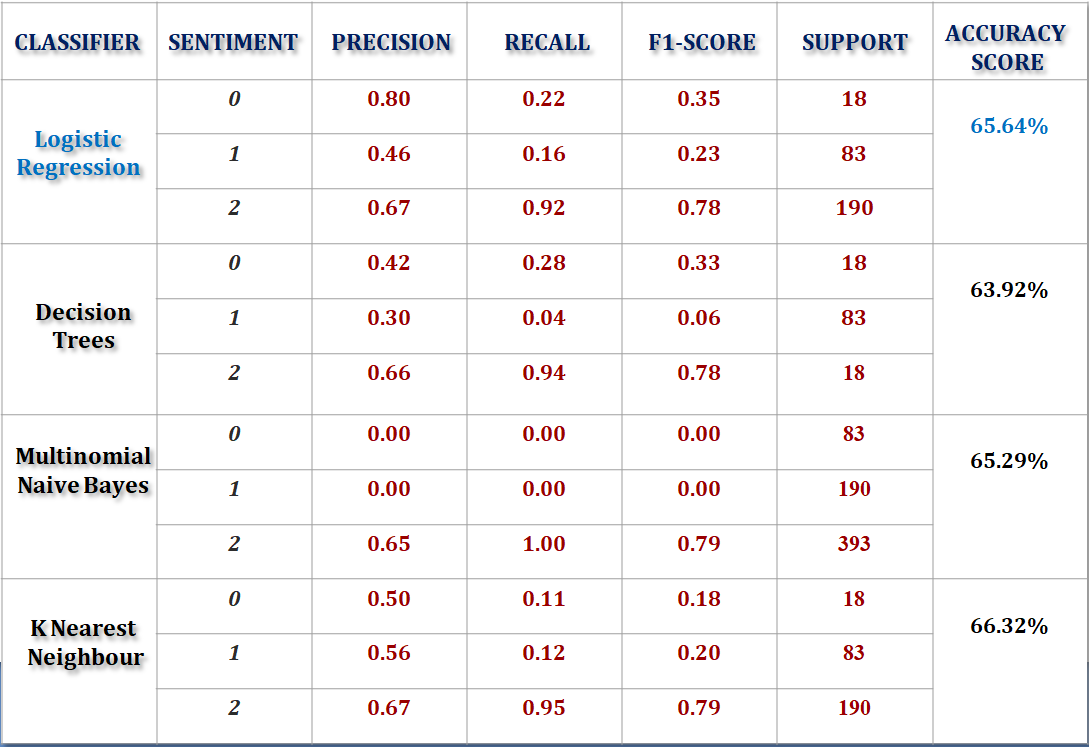
**Tabel 2** – Rezultatele obținute pe baza clasificatorilor utilizați pentru clasificarea automată



# 9.2. Clasificarea manuală

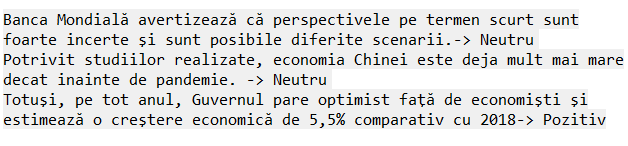
În cadrul acestui proiect am realizat și clasificarea manuală a știrilor financiare iar pentru realizarea acestui lucru s-au clasificat doar 20% din instanțele prezente în baza noastră de date deoarece acest tip de clasificare necesită foarte mult timp pentru a fi efectuată. Pe baza rezultatelor obținute observăm că și în acest caz cel mai bun clasificator este reprezentat de regresia logisitică și, de asemenea dacă comparăm cele două tipuri de clasificări pe baza rezultatelor observăm că avem o acuratețe mai ridicată în privința clasificatorilor aplicați pe știrile financiare unde sentimentele au fost identificate automat.

**Tabel 3** – Rezultatele obținute pe baza clasificatorilor utilizați pentru știrile clasificate manual



# 10. Clasificarea unor noi observații

În cadrul proiectului pe baza celui mai bun clasificator reprezentat de **regresia logistică** ce a fost antrenat pe setul de date considerat vom incerca să clasificăm un număr de 3 noi știri financiare. Astfel că pe baza output-ului putem observa faptul că dintre cele 3 noi știri financiare, clasificatorul ales reușește să clasifice doar două știri în mod corect ceea ce este de înțeles dat fiind faptul că acuratețea acestui clasificator este de numai **76,34%.**



**Figura 7** – Output pentru clasificarea unor noi știri financiare

# 11. Contribuție proiect

Prezentul proiect nu a necesitat neapărat alegerea unui lider iar contribuția fiecărui membru de echipă pe parcurusul realizării acestuia este prezentată pe scurt în continuare.



# Concluzii

Utilitatea realizării unei analize sentiment de clasificare a titlurilor de știri financiare se explică prin faptul că traderii și comercianții caută informații despre companiile de la care vor să cumpere acțiuni pentru tranzacționare pe termen lung și scurt sau să investească într-o anumită companie iar în acest sens o sursă frecventă de informații sunt **știrile financiare** care oferă informații actualizate despre activitățile unei companii, despre veniturile mai bune / mai slabe pe care aceasta se le deține, noi produse și multe altele. În funcție de aceste știri financiare, acești comercianți pot determina un trend, o anumită tendință și pe baza acesteia decid dacă să investească sau nu în ea. Astfel că **identificarea sentimentelor știrilor financiare** îi ajută foarte mult pe traderi în luarea decizilor de a investi într-o companie, de a-și realiza strategii de trading dar mai ales de a lua decizii într-un timp cât mai scurt.

În cadrul acestui proiect s-a urmărit realizarea unei analize sentiment de clasificare a titlurilor de știri financiare cu ajutorul unor clasificatori și, de asemenea realizarea clasificării unor noi știri financiare pe baza celui mai bun clasificator ales. În urma comparării indicatorilor de performanță pentru fiecare clasificator în parte pe baza rezultatelor s-a observat că cel mai bun clasificator este reprezentat de regresia logistică dat fiind faptul că prezintă cea mai ridicată acuratețe a modelului de 76,34% ceea ce înseamnă că reușește să clasifice cel mai bine titlurile de știri financiare.

# Bibliografie

1. <https://www.kaggle.com/ankurzing/sentiment-analysis-for-financial-news>
2. <https://www.kaggle.com/shwethaspdevam/comparison-of-models-for-financial-news>
3. <https://www.researchgate.net/profile/Rob_Schumaker/publication/255604315_Sentiment_Analysis_of_Financial_News_Articles/links/56d49bcb08ae2cd682b93c7c.pdf>
4. <https://arxiv.org/pdf/1811.11008.pdf>
5. <https://www.researchgate.net/publication/338898380_Financial_News_Sentiment_Analysis_Using_Lexicon-_Based_Labelling_and_Machine_Learning-Based_Algorithm>
6. <http://ceur-ws.org/Vol-862/FEOSWp4.pdf>
7. <https://www.kaggle.com/shwethaspdevam/comparison-of-models-for-financial-news>
8. <https://www.kaggle.com/ankurzing/sentiment-analysis-for-financial-news>