Detecció de Notícies Falses

Blanca Piñol Chacon[a,1] **and Maria Muñoz Cabestany**[b,2]

[a] *1666134*
[b] *1668438*

Abstract—En aquest treball s'ha desenvolupat un sistema per a la detecció de notícies falses utilitzant tècniques de processament de text i models de classificació. Es va utilitzar un conjunt de dades de notícies etiquetades com a verídiques o falses, amb l'objectiu de construir un model capaç de predir la veracitat d'una notícia basant-se en les seves característiques. Per ampliar el conjunt de dades, es va realitzar un scraping de 500 notícies verídiques, equilibrant així el dataset. A continuació, es va netejar la informació eliminant incoherències i tractant els valors nuls. A més, es van estàndarditzar les notícies detectant-ne els idiomes i traduint-les quan calia.
Les columnes textuals es van vectoritzar mitjançant els mètodes de Bag-of-Words (BoW) i TF-IDF. Per a cada un d'aquests mètodes, es van provar diferents configuracions, incloent l'ús només del text, només del títol i combinacions del text i el títol amb altres característiques com la presència d'una imatge (variable binària) i el nom de l'autor. Es va utilitzar el model Naive Bayes, concretament els models MultinomialNB, BernoulliNB, ComplementNB i GaussianNB, per a la classificació. Es va realitzar una validació creuada i una cerca d'hiperparàmetres per avaluar el millor rendiment del model
Els resultats van indicar que el millor model es va obtenir utilitzant TF-IDF amb ComplementNB i només el text. Aquesta configuració va ser validada amb un conjunt de dades de prova, aconseguint una bona precisió en la predicció de notícies verídiques i falses.

Contents

1. INTRODUCCIÓ

Avui en dia, que gaudim d'un accés instantani a la informació, el fenòmen de les notícies falses és cada cop més present. Aquest creixement constant de desinformació ens va portar a voler treballar en la detecció automàtica de notícies falses.
Aquest treball aborda aquest problema partint d'un dataset equilibrat gràcies a l'implementació de scraping. L'aborda mitjançant tècniques de processament del llenguatge natural i aprenentatge automàtic. S'han aplicat mètodes com la vectorització de textos amb TF-IDF (Term Frequency-Inverse Document Frequency) i Bag-of-Words, així com la implementació de 4 classificadors diferents basats en el teorema de Naive Bayes i altres tècniques d'aprenentatge supervisat. Aquest enfocament permet analitzar la fiabilitat d'un text basant-se en les seves característiques.
L'objectiu principal és desenvolupar un model que no només sigui capaç de distingir entre notícies falses i reals amb una precissió considerable i amb bona generalització, sinó també entendre les limitacions i possibles millores del model. Les dades d'aquest estudi han estat preprocessades i analitzades per garantir bons resultats.

2. Conjunt de dades i Scraping

El nostre conjunt de dades va ser creat i penjat a Kaggle per Ruchi Bhatia: https://www.kaggle.com/datasets/ruchi798/source-based-news-classification

Per tal d'equilibrar el nombre de notícies verídiques amb les falses, s'ha realitzat Scraping a través de la pàgina web `News API`. Es van recollir 500 notícies actuals sobre diferents temes (tecnologia, política, salut, economia i esports), les quals es van afegir al dataset anterior.

Finalment el nou conjunt de dades va quedar amb 2596 files i 12 columnes.
Les columnes són les següents:

- **author:** Autor de la notícia (object)
- **published:** Any, mes, dia i hora en que s'ha penjat la notícia (object)
- **title:** Títol de la notícia (object)
- **text:** Contingut de la notícia (object)
- **lenguage:** Idioma de la notícia (object)
- **site_url:** Pàgina web de la notícia (object)
- **main_img_url:** Enllanç de la imatge de la notícia (object)
- **type:** Tipo de notícia (object)
- **label:** Indica si la notícia és Real o Falsa (object)
- **title_without_stopwords:** Títol de la notícia sense stopwords (object)
- **text_without_stopwords:** Contingut de la notícia sense stopwords (object)
- **hasImage:** Indica si té imatge o no (float64)

De les 2596 notícies, 1301 són reals i 1294 falses.

2.1. Tractament de nans

Per tal de predir correctament, el nostre conjunt de dades no ha de tenir Nans.

Primer es va mirar quins columnes en tenien. Per això es van convertir a Nans aquelles caselles on no hi havia l'informació adequada, és a dir, on posava directament "NO *característica*", Removed o nan però en format text.

Un cop realitzat el canvi, podem veure els nans reals del dataset:

- **author:** 65
- **title:** 205
- **text:** 65
- **lenguage:** 1
- **site_url:** 1
- **main_img_url:** 509
- **type:** 1
- **label:** 1
- **title_without_stopwords:** 189
- **text_without_stopwords:** 69
- **hasImage:** 1

Una notícia sense text no es pot considerar una notícia, per tant es van eliminar totes aquelles files on la varibale text estigues buida.
La columna del url de les imatges es va suprimir ja que aquests no es podien obrir de manera segura.

Un cop fet això, només quedaven amb nans les varibales author, title, title_without_stopwords i text_without_stopwords
Els Nans del autor es van quedar com No Author, ja que no es va trobar la manera de saber els autors de les notícies que faltaven.
El text_without_stopword es va omplir fent servir la llibreria `ntlk` , la qual elimina les stopwords utilitzant el corresponent text.
Per les notícies que si tenien title_without_stopwords, però no title, es va assignar el mateix a les dues columnes. Per la resta de title buids es va utilitzar una tècnica anomenada Rake(Rapid Automatic Keyword Extraction), que selecciona les paraules clau més importants del text i les combina per formar un títol. Es va fer servir la llibreria `rake_nltk`. Pel title_without_stopwords es va seguir el mateix procediment que al eliminar les stopwords del text.

3. Inconsistències en algunes columnes

Encara que en el dasaset ja no hi hagués Nans, hi havia algunes columnes que no estaven arreglades.

3.1. Published

A l'hora de fer l'Scraping, la data de publicació de les diferents notícies estava expressada en zones horàries diferents i amb un format diferent. Algunes dates estaven expressades en UTC (Coordinated Universal Time), mentre que d'altres utilitzeven altres zones horàries locals, com l'horari d'Europa o d'Amèrica, que tenen desplaçaments respecte a UTC.

Per solucionar aquesta diversitat de formats i zones horàries, es van transformar totes les dades a UTC (per ser universal) i en format ISO 8601 ('AAAA-MM-DDTHH:mm:ss+00:00'). Es va utilitzar la llibreria `deteutil` per realitzar aquesta feina.
Finalment hi van haver dues notícies que no es van poder tranformar perquè eren links i van ser eliminades.

3.2. Lenguage

De primeres el dataset mostrava que hi havia notícies en diferents idiomes: anglès, alemany, francés, espanyol i ignore.

En relació a ignore, com eren poques, es va comprovar que estaven en inglés manualment.
Imprimint unes quantes més, es va detectar que l'idioma no era el correcte en alguns casos. Per tant, per sortir de dubtes, es va utilitzar la llibreria `langdetect` per tal de detectar l'idioma real del text.

Una vegada fet això, es van detectar dos idiomes, l'alemany i l'anglés. Per tenir totes en la meteixa llengua, es van traduir les que estaven en alemany a l'anglés. Per aquest treball es va utilitzar la llibreria del Google Translator, `deep_translator`. Dues notícies no es van podien traduir, així que van ser eliminades.

4. Anàlisi Exploratòria de Dades (EDA)

En aquesta secció es va dur a terme una anàlisi exploratòria per entendre millor el conjunt de dades i identificar possibles relacions entre les característiques.
Abans de començar mostrem que el nostre conjunt de dades està balancejat.

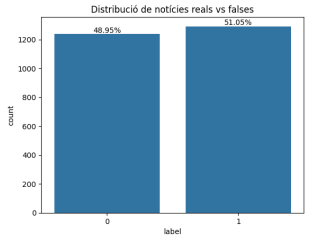

Figure 1. Percentatge de notícies reals i falses

4.1. Freqüència de paraules

Per analitzar les paraules més freqüents, es va utilitzar la llibreria CountVectorizer de `scikit-learn`. Aquest procés va permetre identificar les paraules amb major presència, tant en notícies verídiques com en falses.

4.1.1. Text_without_stopwords

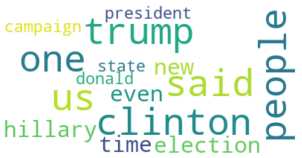

Figure 2. Les 20 paraules més freqüentes

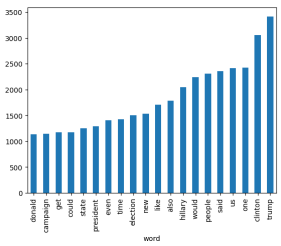

Figure 3. Comptatge de les 20 paraules més freqüents

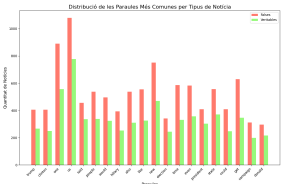

Figure 4. Vericitat de les 20 paraules més freqüents

4.1.2. Title_without_stopwords

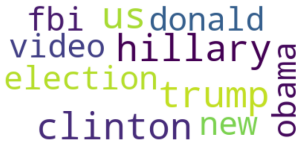

Figure 5. Les 10 paraules més freqüentes

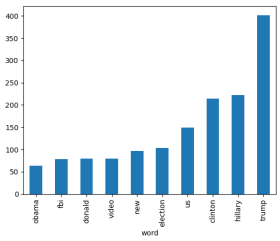

Figure 6. Comptatge de les 10 paraules més freqüents

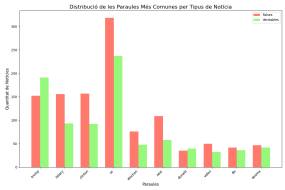

Figure 7. Vericitat de les 10 paraules més freqüents

4.2. Distribució de la característica hasImage

Es va examinar la distribució binària de la columna `hasImage`, que indica si la notícia conté o no una imatge associada.

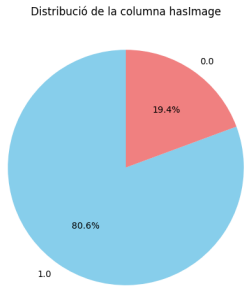

Figure 8. Percentatge de la columna HasImage

4.3. Anàlisi d'autors

Es va detallar la distribució dels autors per identificar quins tenien més notícies associades. Es van comptar les notícies per autor, es van harmonitzar alguns valors inconsistents en els noms dels autors i es va filtrar el conjunt de dades per als 10 autors més freqüents.
Es van fer els gràfics utilitzant la llibreria `seaborn`.

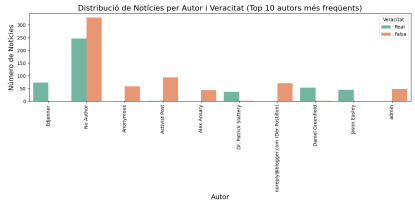

Figure 9. Veracitat dels 10 autors més freqüents

4.4. Anàlisi published

Es va realitzar un anàlisi de la quantitat de notícies publicades cada any i quantes eren reals i quantes falses.

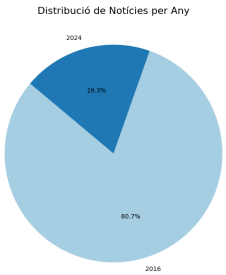

Figure 10. Any de les notícies

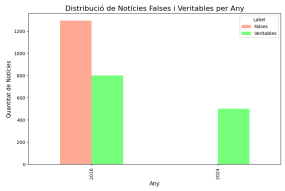

Figure 11. Vericitat de les notícies per any

Es pot veure que la nostra columna published no és una bona representació ja que totes les notícies del dataset principal eren de 2016 i les introduides per l'scraping (totes reals) eren de 2024. Per tant no ens aporta informació rellevant.

4.5. Anàlisi type

Es va analitzar els diferents tipus de notícies.

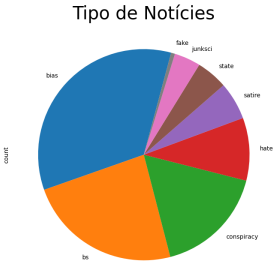

Figure 12. Tipus de les notícies

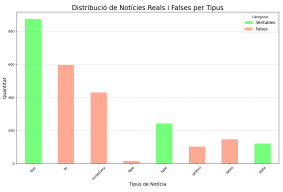

Figure 13. Vericitat de les notícies per tipus

Es pot veure com cap tipus té notícies reals i falses alhora. Per tant, la columna tipus no ens donarà informació útil.

4.6. Eliminació de columnes no rellevants

Per simplificar el conjunt de dades i centrar-nos en les característiques més importants, es van eliminar les següents columnes:

- **published:** Data i hora de publicació. Tot i estar normalitzada a un format coherent, no aportava informació rellevant per als models.
- **site_url:** URL del lloc web de la notícia. La majoria de links no eren segurs o no estaven disponibles, per tant, aquesta columna no contenia informació de la qual en poguessim treure profit.
- **language:** Idioma de la notícia. Després de deixar totes les notícies en anglès, aquesta columna no és necessària.
- **type:** Tipus de notícia. Aquesta columna no contenia informació realista, per tant, es va eliminar.
- **title:** Títol de la notícia. Aquesta columna va estar substituïda per `title_without_stopwords`, que elimina les paraules innecessàries i conserva el contingut essencial.
- **text:** Contingut complet de la notícia. Igual que en el cas del títol, aquesta columna es va substituir per `text_without_stopwords`, que conté les paraules significatives.

5. Vectorització i Classificació amb Bag-of-Words

L'objectiu en aquest cas era transformar les dades textuals en representacions numèriques per tal d'entrenar models de classificació. Per a això, es va utilitzar la tècnica Bag-of-Words (BoW), que permet convertir el text en una matriu de nombres que depenen de la freqüència de les paraules. Aquesta tècnica es va combinar amb diversos models de classificació Naive Bayes per avaluar el seu rendiment.

5.1. Vectorització amb Bag-of-Words

Per dur a terme la vectorització, es va utilitzar la la llibreria CountVectorizer de `scikit-learn`, seleccionant les 5000 paraules més freqüents per limitar la dimensió de la matriu i evitar problemes de sobreajustament.
Per garantir que els resultats fossin fiables, es va dividir el conjunt de dades en un 80% per entrenament i un 20% per test, mantenint la proporció original entre notícies verídiques i falses mitjançant estratificació.

5.2. Cross-Validation

Un cop realitzada la vectorització, es van entrenar i avaluar diversos models de Naive Bayes:

- **MultinomialNB:** Model dissenyat per treballar amb dades discretes com son les generades per BoW.
- **BernoulliNB:** Adaptat per a característiques binàries, útil per detectar presència o absència de termes.
- **ComplementNB:** Una variant de MultinomialNB, especialment eficient amb conjunts de dades desequilibrats.
- **GaussianNB:** Model pensat per treballar amb dades contínues, utilitzat en aquest cas després de convertir la matriu BoW en un array.

Per mesurar el rendiment dels models, es va utilitzar el F1-score mitjançant validació creuada amb cinc particions. Es va escollir aquesta mètrica ja que voliem tenir controlat tant els falsos negatius com els falsos positius.

5.3. Diferents combinacions de les característiques

Els quatre models esmentats anteriorment es van implementar per cada una d'aquestes opcions:

- text_without_stopwords
- title_without_stopwords
- text_without_stopwords + hasImage
- title_without_stopwords + hasImage
- title_without_stopwords + author

5.4. Optimització d'hiperparàmetres

Per millorar el rendiment dels models, es va dur a terme una cerca dels millors hiperparàmetres, descartant els models que anteriorment havien resultat en un F1-score més baix amb diferència respecte la resta. Els millors resultats es van obtenir amb la combinació de característiques autor i títol amb:

- **MultinomialNB:** `alpha=1.0`, F1-score: 0.8291
- **ComplementNB:** `alpha=0.5`, `norm=True`, amb `norm=True`, F1-score: 0.8231

Seguit de la combinació de títol i imatge amb:

- **BernoulliNB:** `alpha=6.0`, `binarize=0.0`, F1-score: 0.7389

5.5. Resultats finals i comparativa de configuracions

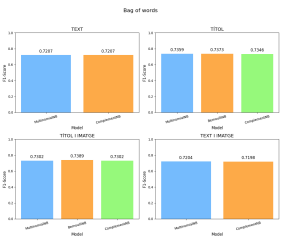

Figure 14. Resultats models amb BoW

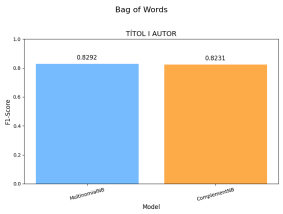

Figure 15. Resultats de títol+autor amb BoW

Per tant l'anàlisi mostra que la configuració òptima es va aconseguir utilitzant la combinació de la característica `tittle_without_stopwords` i `author`, obtenint un F1-score de 0.8231 amb el ComplementNB i un F1-score de 0.8291 amb el MultinomialNB.
Seguidament, el millor resultat ve donat pel model BernoullitNB utilitzant el title_without_stopwords i hasImage amb un F1-score d 0.7389 .

6. Vectorització i Classificació amb TD-IDF

L'objectiu d'aquesta etapa era transformar les dades textuals en representacions numèriques utilitzant la tècnica TF-IDF (*Term Frequency-Inverse Document Frequency*), que permet assignar un pes a cada paraula en funció de la seva freqüència en un document i la seva rellevància. És especialment útil per reflexar la importància relativa de les paraules i minimitzar l'impacte de termes molt comuns però poc significatius. Es van utilitzar les implementacions de TF-IDF disponibles a la llibreria `scikit-learn`.

6.1. Vectorització amb TF-IDF

Per dur a terme la vectorització, es va utilitzar la llibreria TfidfVectorizer de `scikit-learn`. Per garantir que els resultats fossin fiables, de nou, es va dividir el conjunt de dades en un 80% per entrenament i un 20% per test.

6.2. Cross-Validation

Un cop realitzada la vectorització, es van entrenar i avaluar els mateixos models de Naive Bayes que amb el BoW. Per mesurar el rendiment dels models , també es va seguir el mateix procediment.
Les diferents combinacions de les característiques també van ser les mateixes que amb el BoW.

6.3. Optimització d'hiperparàmetres

Per millorar el rendiment dels models, es va dur a terme una cerca dels millors hiperparàmetres, descartant de nou en la cerca els models que tenien un F1-score més baix amb diferència respecte la resta. Els millors resultats es van obtenir amb la combinació característiques autor i títol amb:

- **MultinomialNB:** `alpha=0.5`, F1-score: 0.8558
- **ComplementNB:** `alpha=0.5`, `norm=True`, amb `norm=True`, F1-score: 0.8589

Seguit de l'opció que només considera el text de la notícia:

- **MultinomialNB:** `alpha=12.0`, F1-score: 0.8033
- **ComplementNB:** `alpha=2.0`, `norm=True`, amb `norm=True`, F1-score: 0.8061

6.4. Resultats finals i comparativa de configuracions

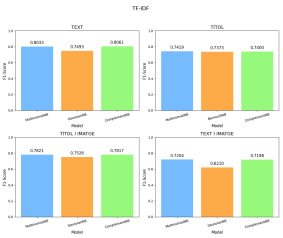

Figure 16. Resultats models amb TF-IDF

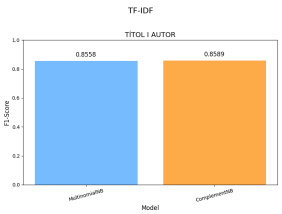

Figure 17. Resultats de títol+autor amb TF-IDF

Per tant l'anàlisi mostra que la configuració òptima es va aconseguir utilitzant la combinació de la característica `tittle_without_stopwords` i `author`, obtenint un F1-score de 0.8589 amb el ComplementNB i un F1-score de 0.8558 amb el MultinomialNB.
Seguidament, el millor resultat ve donat pel model ComplementNB utilitzant el text_without_stopwords amb un F1-score d 0.8061.

7. Anàlisi final

En l'anàlisi, es va observar que incloure l'autor com a característica inicialment millorava les mètriques de classificació. No obstant això, aquesta millora no era consistent. Una part significativa del nostre conjunt de dades conté articles amb autors anònims o sense informació específica sobre l'autor. Els que no són anònims, són, en gran majoria, un tipus d'autor concret. Això fa que aquesta característica no sigui representativa ni fiable per al nostre propòsit i que no contribueixi a un model que generalitzi adequadament. Per tant, es va acabar descartant.

El millor rendiment es va obtenir utilitzant només el text dels articles i transformant-lo amb el mètode TF-IDF. D'entre els models provats, el **MultinomialNB** i el **ComplementNB** van obtenir els millors resultats. Finalment, el **ComplementNB** es va seleccionar com el model òptim gràcies al seu lleuger avantatge en les mètriques finals. Aquest model es va entrenar amb els paràmetres `alpha=2.0` i `norm=True`.

7.1. Resultats del Model ComplementNB

- **Accuracy (Test):** 0.75
- **F1 Score (Test):** 0.78

Després d'aplicar el ComplementNB a les dades, es va obtenir una accuracy de 0.75, la qual cosa indica que el model classifica correctament el 75% de les mostres. Pel que fa al F1-score vam obtenir un valor de 0.78, el que reflecteix un bon equilibri entre la precisió i el recall.

La matriu de confusió mostra la distribució de prediccions encertades i errònies (en percentatges) per a les classes "REAL" i "FALSA":

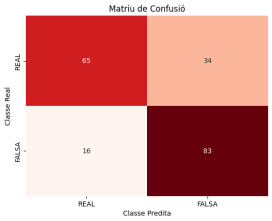

Figure 18. Matriu de confusió

Aquesta informació ens confirma que el model pot classificar notícies com "REAL" i "FALSA" amb una precisió considerable. Tot i així, encara hi ha marge de millora en la classe "REAL", on té lloc una tendència a predir falsos positius.

7.2. Corbes ROC

Per avaluar el rendiment global del model es va utilitzar la corba ROC, ja que el dataset està equilibrat.

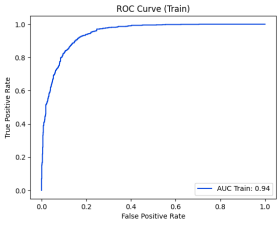

Figure 19. Corba ROC pel train

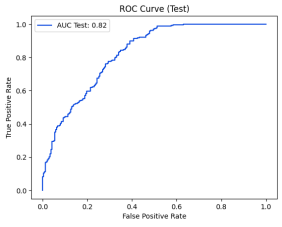

Figure 20. Corba ROC pel test

Es veu com els resultats són bons, tenien un AUC de 0.94 al train i un AUC de 0.82 al test.

8. Cas d'ús

Per il·lustrar l'ús del model de classificació Complement Naive Bayes, es va realitzar un anàlisi de notícies reals i falses utilitzant el mètode TF-IDF per a la vectorització del text. El procés inclou l'obtenció de textos a partir de dues notícies específiques, una de verídica i una altra de falsa, i la seva posterior classificació mitjançant el model entrenat.

Per aquest cas d'ús, es van seleccionar dues notícies d'exemple: una notícia verídica de la BBC, una empresa de transmissió de notícies molt coneguda, i una notícia falsa de The Onion, una empresa que publica articles satírics. Els enllaços corresponents són els següents:

- **Notícia verídica:** https://www.bbc.com/news/articles/c62l5zdv7zko
- **Notícia falsa:** https://theonion.com/nations-mumblers-march-on-washington-demanding-something-or-other/

Per tal d'analitzar-les, primer es van obtenir els textos de les notícies utilitzant la biblioteca `newspaper3k`. Un cop obtinguts, es va realitzar un procés de neteja eliminant les stopwords.

8.1. Resultats de la Classificació

Amb el model Naive Bayes entrenat, es va aplicar la predicció a les dues notícies. Els resultats van ser els següents:

- **Notícia verídica de BBC:** La predicció va ser **Verídica**.
- **Notícia falsa de The Onion:** La predicció va ser **Falsa**.

9. CONCLUSIONS FINALS

En aquest treball s'ha desenvolupat un sistema de detecció de notícies falses mitjançant tècniques de vectorització textual, com Bag-of-Words i TF-IDF, combinades amb models de classificació basats en Naive Bayes. Els resultats obtinguts mostren que, tot i que ambdues tècniques aporten beneficis, TF-IDF genera una millora significativa en el rendiment general, especialment quan s'utilitza conjuntament amb el model ComplementNB i el contingut del text (`text_without_stopwords`).

El model ComplementNB es basa en un principi que s'ajusta millor a la distribució de les dades quan aquestes no segueixen una distribució multinomial perfecta. Aquesta propietat el fa més eficaç en tractar dades disperses, com les generades per la tècnica de TF-IDF.

La configuració utilitzada per entrenar el model va ser un valor d'alpha de 2.5 i la normalització activada (`norm=True`). Els resultats obtinguts amb el conjunt de prova han estat molt bons, amb un accuracy de 0.75 i un F1-score de 0.78. A més, les corbes ROC mostren un AUC de 0.94 per al conjunt d'entrenament i un AUC de 0.82 per al conjunt de prova, indicant que, tot i una lleugera disminució en el rendiment sobre el conjunt de prova, el model encara obté uns resultats sòlids.

Aquesta conclusió es valida amb els resultats obtinguts en el cas d'ús, en què, mitjançant l'aplicació del model a dues notícies, es comprova que les prediccions són correctes. Els bons resultats obtinguts en aquest experiment demostren que el sistema és efectiu per a la classificació de notícies falses.

9.1. Millores

Les possibles millores que es podrien implementar en el futur són les següents:

- Tot i que els resultats són força bons, s'ha observat que afegir la informació de l'autor millora els resultats. En aquest cas, s'ha descartat aquesta opció perquè les dades de la columna author no són prou representatives. Una millora futura podria consistir a treballar amb notícies amb més informació sobre l'autor, és a dir, notícies on aquest estigués més ben definit, per així analitzar si realment aquesta informació té una influència positiva en la predicció.
- Si els URLs de les imatges haguessin estat vàlids, aquesta informació podria haver-se inclòs com una altra característica per millorar la predicció.
- També es podrien haver utilitzat altres models de classificació diferents del *Naive Bayes*, com ara *Logistic Regression*, per comprovar si ofereixen un millor rendiment.
- Hem obervat que aquest model és més bo per detectar notícies falses que no pas verídiques, un aspecte amb el que es podria treballar.