UNIVERSITÀ DEGLI STUDI DI

MODENA E REGGIO EMILIA

Laurea in Informatica

Natural Language Processing e classificazione: modelli predittivi per l'analisi di una comunicazione efficace in ambito Cultural Heritage.

Candidato
Enrico Fiorini

Relatori
Prof. Riccardo Martoglia
Prof.ssa Manuela Montangero

Anno Accademico 2020/2021

AMBITI DELLA TESI

Comunicazione sui social media

- Tweet di enti operanti in settore cultural heritage
- Miglioramento dell'efficacia comunicativa
- Social media managment

Analisi previosionale e Machine Learning

- Natural Language Processing
- Previsione valori futuri
- Metodologie statistiche e di machine learning

OBIETTIVI DELLA TESI

A

· Studio degli strumenti e delle metodologie più opportune per eseguire analisi previsionali sull'apprezzabilità di un tweet postato da un social media manager

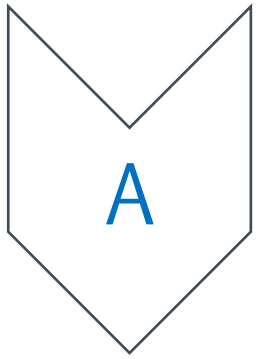
B

· Implementazione delle metodologie usate in un modello predittivo in grado di gestire grandi quantità di tweet e darne una classificazione

C

· Valutazione dell'accuratezza e dell'impatto dell'analisi quale elemento migliorativo ai fini dell'analisi previsionale di apprezzabilità di un tweet

SEZIONE A



Studio degli strumenti e delle metodologie

STRUMENTI UTILIZZATI

- Linguaggio di scripting orientato agli oggetti

- Libreria Python per la manipolazione e l'analisi di dati

- Libreria Python per l'utilizzo di algoritmi di machine learning

STRUMENTI E LIBRERIE II

- Libreria Python per l'utilizzo di algoritmi di machine learning

- Libreria Python per l'utilizzo di algoritmi di named entity recognition

PROBLEMA INIZIALE

I dati di ogni gruppo da uno a cinque - da qui in poi indicati con *G* – vengono mantenuti su file testuali con le seguente struttura:

- N° Retweet
- Fascia oraria
- N° Like
- Tweet

Mediante questi dati, il modello preesistente classifica un tweet come potenzialmente gradito (positivo) o non gradito (negativo) da parte dei followers. L'obiettivo del presente lavoro è migliorare l'accuratezza e il numero di feature coinvolte in fase previsionale.

FEATURES INIZIALI

Totale Retweet

Totale Like

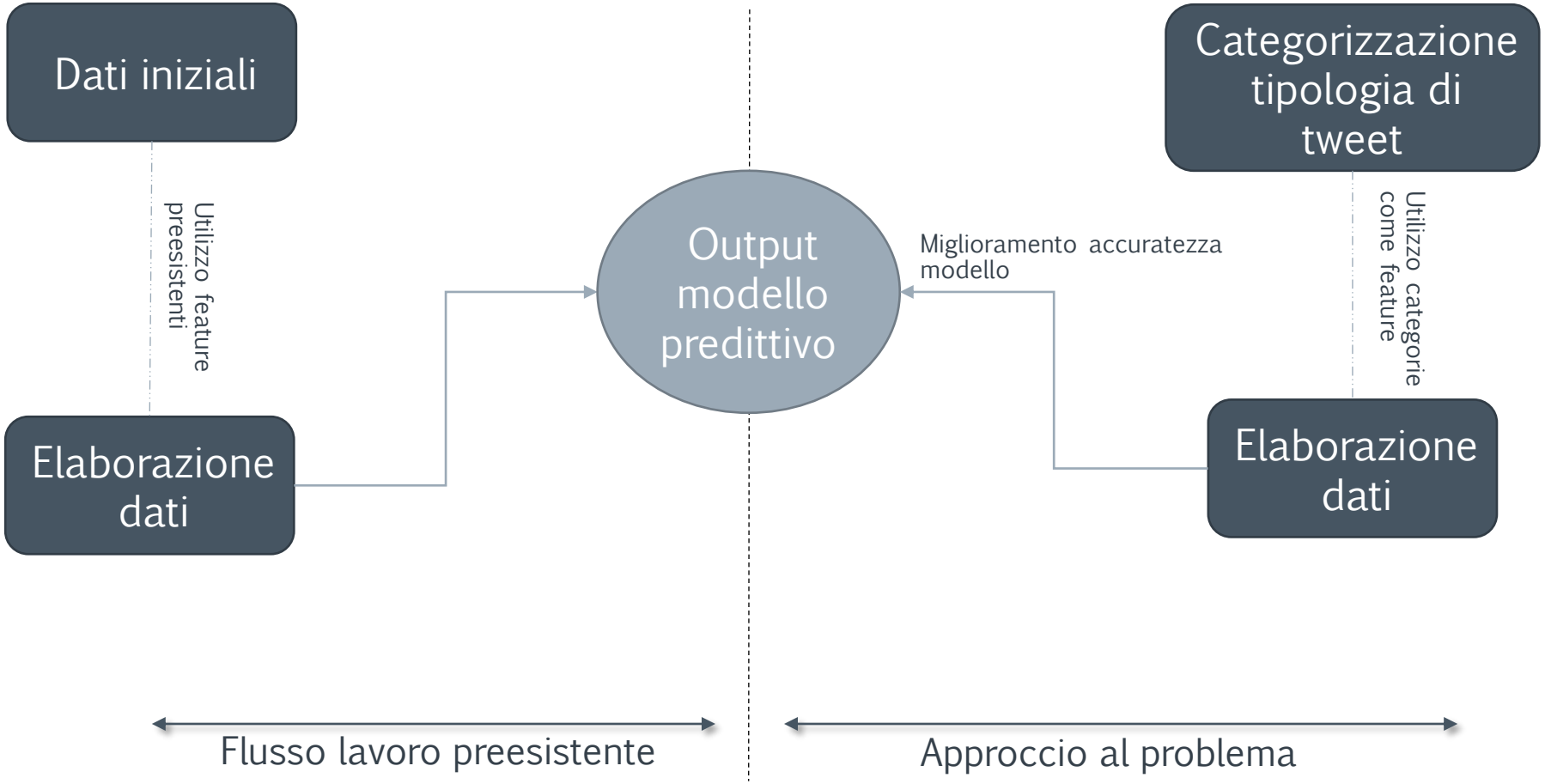
Totale Risposte

Categorie di tweet

Features preesistenti

Feature da introdurre

PROGETTAZIONE



PUNTO DI PARTENZA

Classificatore	Accuratezza G1	Accuratezza G2	Accuratezza G3	Accuratezza G4	Accuratezza G5
XGBClassifier	85.78%	92.60%	79.90%	74.97%	92.23%

Da << Marco Furini, Federica Mandreoli, Riccardo Martoglia, and Manuela Montangero. 2021. A Predictive Method to Improve the Effectiveness of Twitter Communication in a Cultural Heritage Scenario, ACM Journal on Computing and Cultural Heritage (JOCCH), 2022. >> si ottengono i risultati presentati nella tabella sottoprastante relativamente al classificatore XGBClassifier.

PUNTO DI PARTENZA II

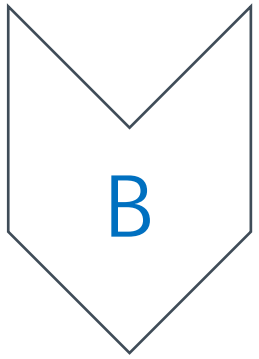
Criteri di raggruppamento musei:

- G1: musei con almeno tre milioni di followers
- G2: musei con più di un milione di followers
- G3: musei con più di 400.000 followers
- G4: musei con più di 200.000 followers
- G5: musei italiani

Data la mancanza di un numero cospicui di tweet in lingua inglese, i gruppi G4 e G5 non verranno presi in considerazione.

I dati dei gruppi sono organizzati su file testuali, contenenti metadati e tweet dell'ente.

SEZIONE B



Dettagli implementativi

CREAZIONE MANUALE DEL DATASET

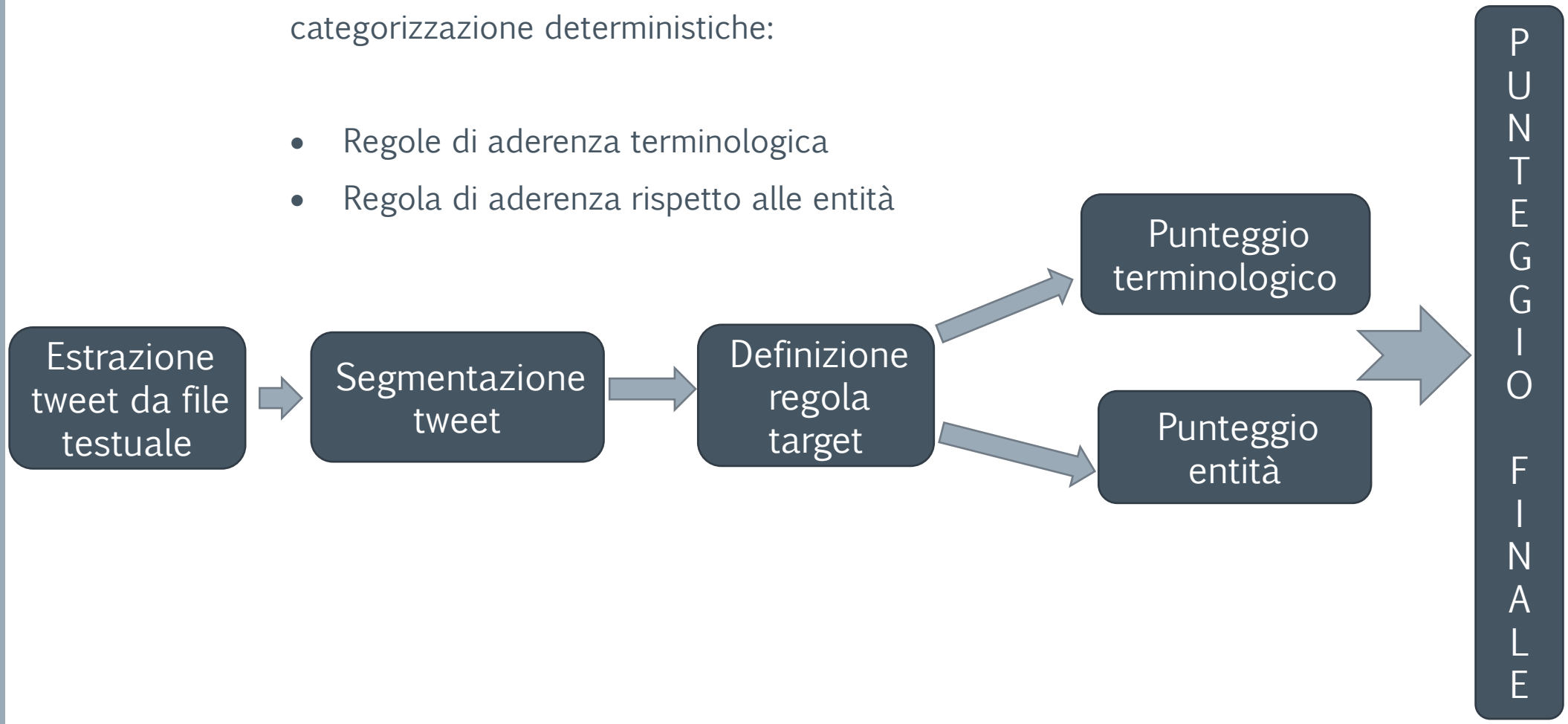
Al termine di un' analisi manuale dei dati iniziali, sono state stilate manualmente le seguenti categorie di classi:

- Artworks
- Festivities (inteso come messaggi di auguri per feste e ricorrenze)
- History (celebrazione di eventi storici, o spiegazione di eventi storici correlati ad opere d'arte esposte)
- #OnThisDay
- Promotions
- VIP & CIT (Celebrazione di personaggi storici, frasi a loro riconducibili, motti storici)

CREAZIONE AUTOMATIZZATA DEL DATASET

Per ognuna delle classi di definite sono state create delle regole di categorizzazione deterministiche:

- Regole di aderenza terminologica
- Regola di aderenza rispetto alle entità



INSERIMENTO AUTOMATICO DI UN TWEET NEL DATASET

Per le regole di aderenza terminologica, un tweet viene valutato secondo:

sia D l'insieme delle keywords per ogni categoria

$$hits(token) = \begin{cases} 1 \ se \ token \ \in D \\ 0 \ se \ token \ \notin D \end{cases}$$

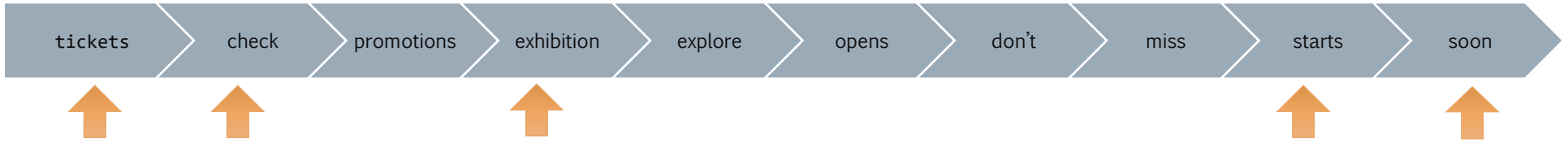
$$Term \ \text{score} = \frac{MAX(\sum_{i=1}^{n} hits(tokenize(tweet), categories[i]))}{\text{n}}$$

$$PUNTEGGIO\ TOTALE = PUNTEGGIO\ ENTITÀ + PUNTEGGIO\ TERMINOLOGICO$$

Mediante l'iterazione su tutte le categorie, si otterrà un punteggio massimo, corrispondente alla categoria individuata per il tweet in input.

ESEMPIO DI INSERIMENTO AUTOMATIZZATO

Si supponga per il seguente esempio come categoria target, la categoria *Promotions*, e la seguente lista di keywords terminologiche:



Per il tweet «*Check our tickets promotion. Exhibition starts soon!*» si avrebbero 5 hit.

Data la formula precedente riportata, a seguito della segmentazione del tweet, esso riporterebbe un punteggio terminologico:

$$\frac{numero\ hits\ per\ regola}{dimensione\ lista\ riferimento} = \frac{5}{10}$$

Iterando per ogni categoria, e tenendo traccia del massimo punteggio ottenuto, si ottiene la categoria a cui legare il tweet da inserire nel dataset.

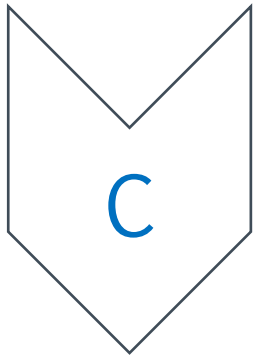
DATASET OTTENUTO

Al termine dell'automazione, si ottiene un dataset di 5637 tweet

Dimensione iniziale del dataset contenente i soli tweet	6791
Totale tweet categorizzati secondo le regole determinate	5637
Totale tweet non categorizzati secondo le regole determinate	1154

Percentuale tweet categorizzati	83%
Percentuale tweet non categorizzati	17%

SEZIONE C



Dettagli implementativi

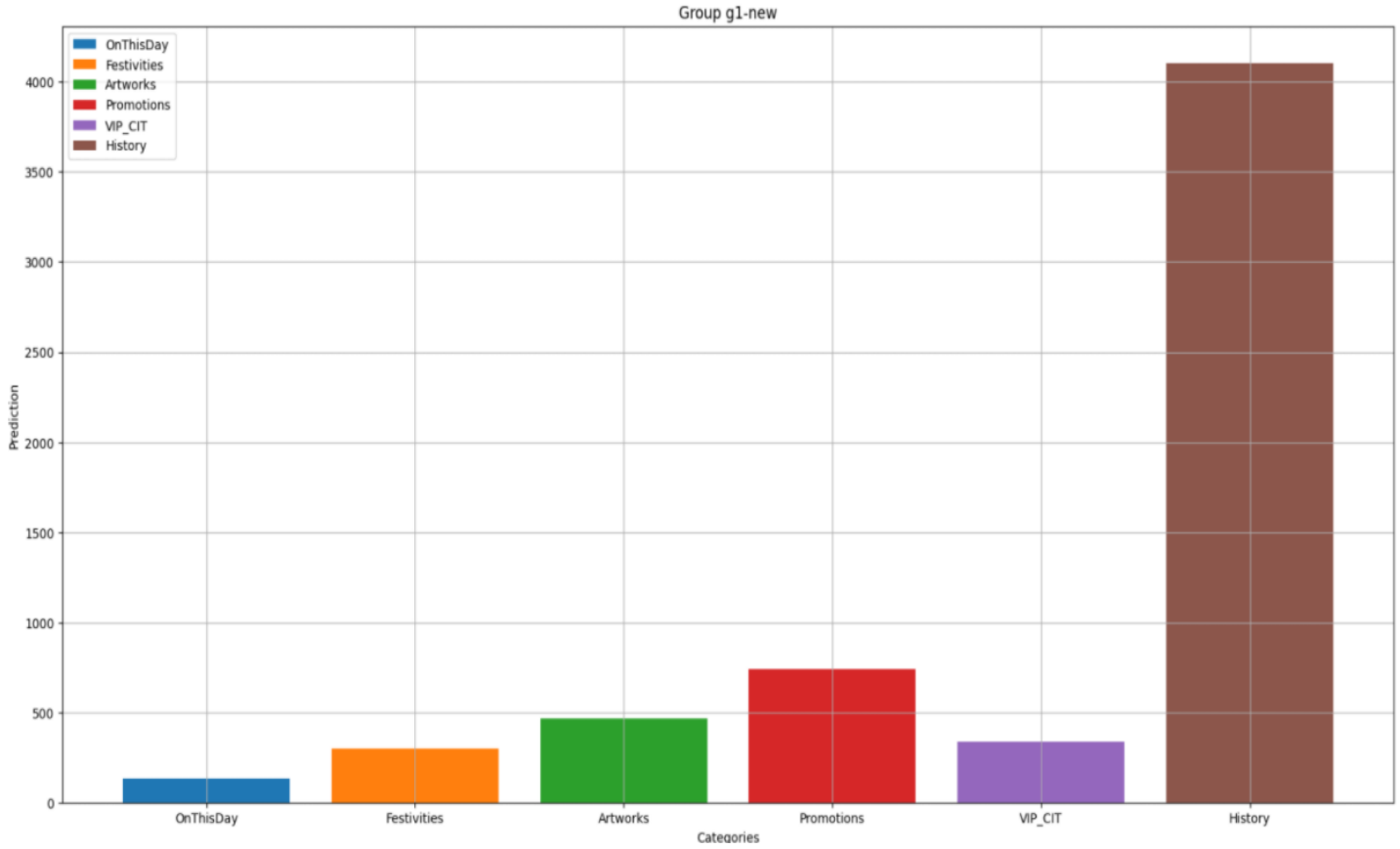
ACCURATEZZA CLASSIFICATORI

Utilizzando il dataset automaticamente costruito, per i seguenti classificatori otteniamo i presenti punteggi di accuratezza in riferimento alle classi di tweet individuate:

Classifier	Accuracy
RandomForestClassifier	83%
KNN Classifier	58%
LogisticRegression Classifier	88%
XGBoost	86%
MultinomialNaiveBayes Classifier	68%

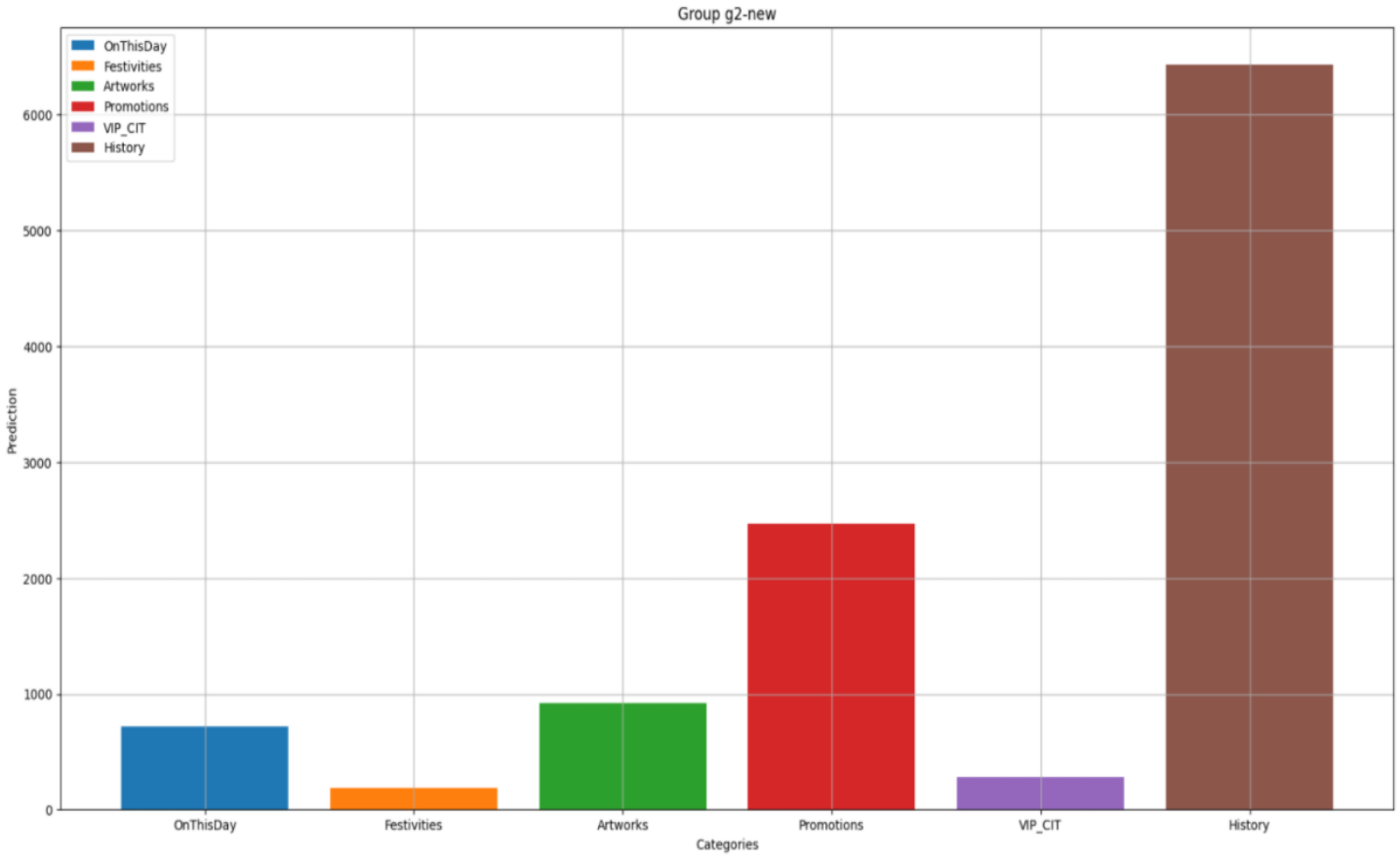
ANALISI GRUPPO G1

Utilizzando il classificatore XGBClassifier con il dataset costruito, su *tutti* i dati del gruppo G1, otteniamo la seguente composizione



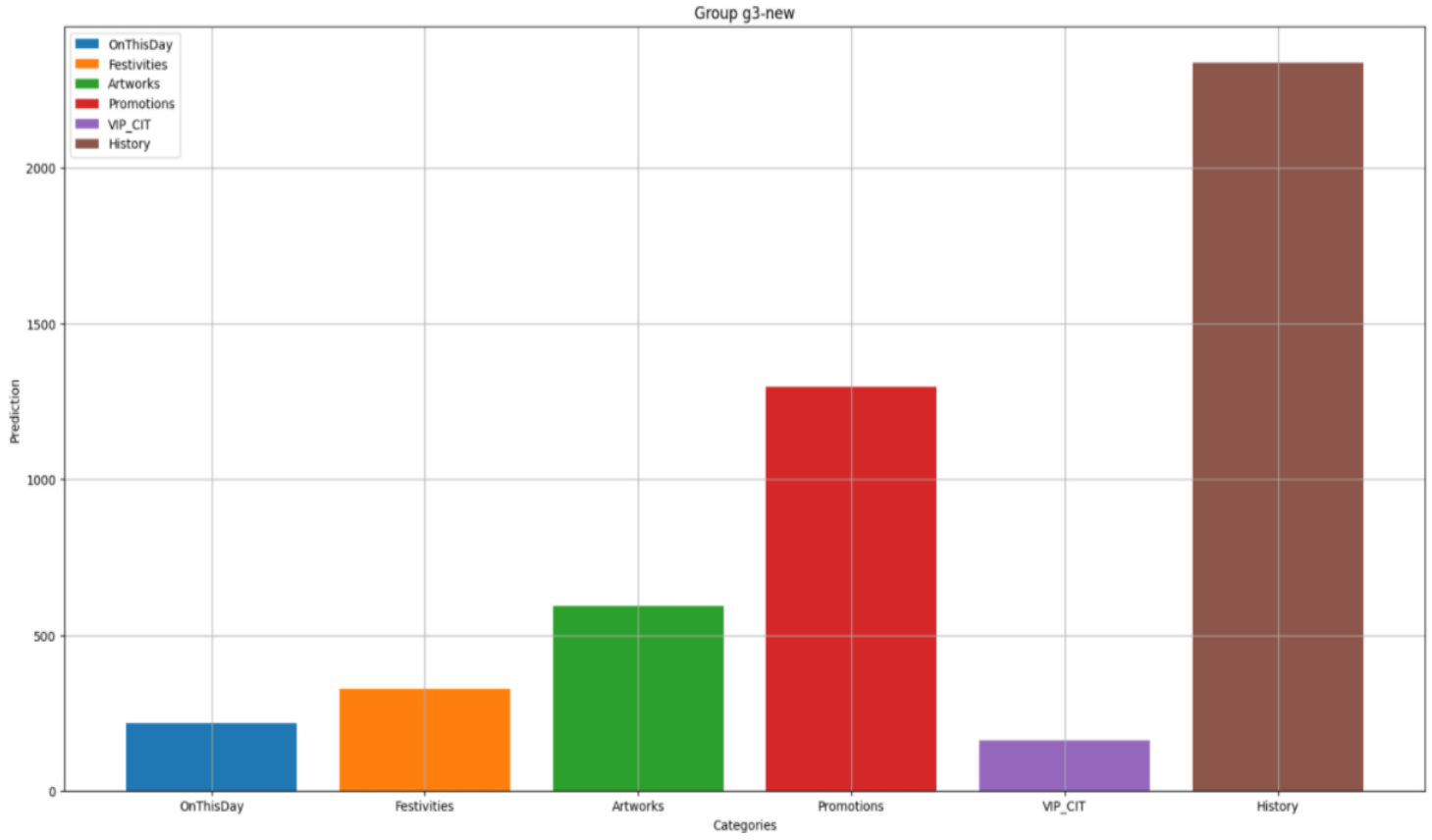
ANALISI GRUPPO G2

Utilizzando il classificatore XGBClassifier con il dataset costruito, su *tutti* i dati del gruppo G2, otteniamo la seguente composizione



ANALISI GRUPPO G3

Utilizzando il classificatore XGBClassifier con il dataset costruito, su *tutti* i dati del gruppo G3, otteniamo la seguente composizione



UTILIZZO DELLA CLASSIFICAZIONE COME FEATURE

Mediante l'utilizzo delle classi di tweet come feature di input del classificatore, otteniamo un miglioramento dell'accuratezza:

Gruppo	Accuratezza % Pre-Feature	Accuratezza % Post-Feature	Incremento efficacia predittiva in %
G1	85.78	87.63	2.16
G2	92.60	92.65	0.05
G3	79.90	81.29	1.74

CONCLUSIONI

Al termine del presente lavoro, si può concludere che:

- Il processo di automazione di creazione del dataset faciliti e migliori il modello predittivo

- L'introduzione della categorizzazione di tweet come feature migliora l'efficacia predittiva e l'accuratezza del modello preesistente.

SEZIONE FINALE

Grazie per l’attenzione