Ferru Daniele Stefano 65036
Ibba Federico 65065

Relazione finale progetto Are 2

Sommario

1. Comanda del progetto 2
2. Struttura dei dataset 2
3. Fasi del progetto 3
3.1 Creazione delle feature 3
3.2 Fase di preprocessing 3
3.3 Creazione pipe per estrapolazione feature 4
3.4 Classificazione 4
3.5 Scelta del Test Set 4
4. Score 4
5. Risultati 5
5.1 Task B 5
5.2 Task C 5
6. Possibili migliorie 5

Ferru Daniele Stefano 65036
Ibba Federico 65065

1. Comanda del progetto

Il progetto deriva da una challenge, il Semeval, che si tiene annualmente. Il task da noi scelto è il task 4 (**Sentiment Analysis**), che era composto da 5 subtask:

A. **Message Polarity Classification**: Given a message, classify whether the message is of positive, negative, or neutral sentiment
B. **Topic-Based Message Polarity Classification**: Given a message and a topic, classify the message on
    a. *two-point scale*: positive or negative sentiment towards that topic
C. **Topic-Based Message Polarity Classification**: Given a message and a topic, classify the message on
    a. *five-point scale*: sentiment conveyed by that tweet towards the topic on a five-point scale
D. **Tweet quantification**: Given a set of tweets about a given topic, estimate the distribution of the tweets across
    a. *two-point scale*: the "Positive" and "Negative" classes
E. **Tweet quantification**: Given a set of tweets about a given topic, estimate the distribution of the tweets across
    a. *five-point scale*: the five classes of a five-point scale.

I subtask da noi scelti sono stati i subtask B e C.

2. Struttura dei dataset

Per ogni task è stato fornito un dataset di train e un dataset di test. Entrambi sono strutturati allo stesso modo con una leggera differenza:

- **Classification task B**: positive, negative, neutral
- **Classification task C**: 2, 1, 0, -1, -2

Di seguito è possibile vedere com'è strutturato il dataset.

Tweet id	Topic	Tweet classification	Tweet text
522712800595300352	aaron rodgers	neutral	I just cut a 25 second audio clip of Aaron Rodgers talking about Jordy Nelson's grandma's pies. Happy Thursday.
523065089977757696	aaron rodgers	negative	@Espngreeny I'm a Fins fan, it's Friday, and Aaron Rodgers is still giving me nightmares 5 days later. I wished it was a blowout.
522477110049644545	aaron rodgers	positive	Aaron Rodgers is really catching shit for the fake spike Sunday night.. Wtf. It worked like magic. People just wanna complain about the L.
522551832476790784	aaron rodgers	neutral	If you think the Browns should or will trade Manziel you're an idiot. Aaron Rodgers sat behind Favre for multiple years.
522887492333084674	aaron rodgers	neutral	Green Bay Packers: Five keys to defeating the Panthers in week seven: Aaron Rodgers On Sunday, ... http://t.co/anCHQjSLh9 #NFL #Packers

Il dataset è stato scaricato utilizzando un servizio di download dei tweet messo a disposizione da Semeval. Alcuni tweet erano "Not Available", quindi sono stati esclusi dal dataset.

Le dimensioni finali del dataset sono quindi risultate le seguenti:

Train Set Task B	Test Set Task B	Train Set Task C	Test Set Task C
15501	4694	23776	11811

3. Fasi del progetto

3.1 Creazione delle feature

La creazione delle feature ha previsto una fase in cui sono state estrapolate informazioni aggiuntive, per ogni tweet, da un servizio IBM di nome Watson. Tramite questo servizio, per ogni tweet, sono stati estrapolati:

- **Categorie**
- **Concetti**

Tutto ciò è stato fatto con l'obiettivo di migliorare l'accuratezza dei classificatori in fase di predizione. In questo modo, sono stati presi i file .tsv contenenti i tweet e sono state aggiunte, con l'utilizzo di un piccolo framework scritto in Java, una colonna per le categorie e una colonna per le feature ottenute da Watson.

Il file finale era così composto alla fine:

tweet_id	topic	tweet_text	class	categories	concepts
104319460143415296	#dexter	Oct. 2 is …	1	/finance/investing/stocks; /science/social science/history/prehistory; /travel/transports/air travel/airplanes	The Return
103641916586999808	#dexter	There is only…	1	/technology and computing/consumer electronics/tv and video equipment/video players and recorders/blu-ray players and recorders;/technology and computing/hardware/computer components/disks;/technology and computing/hardware/computer components/memory/portable	
103277622636716032	#dexter	#3words Tonight's the night #dexter @SHO_Dexter	1	/art and entertainment/music;/art and entertainment/theatre;/business and industrial	Max Beckmann
106222697310208000	#dexter	#ICanHonestlySay #Dexter is such a good show	1	/technology and computing/consumer electronics/tv and video equipment/dvrs and set-top boxes;/sports/boxing;/art and entertainment/movies and tv/movies	

Come si può notare, non sempre sono stati ottenuti concetti o categorie per il tweet fornito.

3.2 Fase di preprocessing

Nella fase di preprocessing sono stati eseguiti sui dati:

- Lower case
- Rimozione di caratteri speciali
- Rimozione di numeri

In aggiunta, come previsto in precedenza, ogni tweet è stato combinato con le sue categorie e i suoi concetti, creando due nuove colonne nel dataframe: **tweet_con** e **tweet_cat**.

3.3 Creazione pipe per estrapolazione feature

Le fasi che deve attraversare ogni tweet sono:

- **Tokenizing**
- **Rimozione delle Stop Word**
- **Count Vectorizing**
- **TF-IDF**

Per fare ciò è stata utilizzata una Pipeline. Ne sono state create due, una per la trasformazione di **tweet_cat** e una per la strasformazione di **tweet_con**.

3.4 Classificazione

La classificazione è diversa in base al task, in quanto il task B necessitava di un classificatore binario, mentre il task C, avendo 5 diverse classi, necessitava di un classificatore multiclasse.

Sono stati così scelti:

- **LinearRegression**: predizione binaria task B
- **DecisionTree**: predizione multiclasse task C

È stata effettuata una classificazione per ogni topic, effettuando poi la media degli score di ogni topic.

3.5 Scelta del Test Set

Dovendo effettuare una predizione per topic, ci si sarebbe aspettati di trovare nell'insieme di test gli stessi topic nell'insieme di train ma in questo caso non è così.

Per ovviare a ciò, si è scelta come strategia quella di cercare nell'insieme di test i tweet più simili al topic da classificare. Per trovare i tweet più simili sono state utilizzate le categorie: i tweet con più categorie in comune a quel topic sono stati presi come test in una percentuale omogenea ad un test_size deciso a priori in precedenza.

4. Score

Gli score richiesti dalla challenge sono diversi in base al subtask:

- **Task B**: richiedeva tre diverse misure, in ordine di importanza:
  - $AvgRec = \frac{1}{2}(R^P + R^N)$
  - $Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}$
  - $F1\ Score = 2\ \frac{precision*recall}{precision+recall}$
- **Task C**: richiedeva due diverse misure, in ordine di importanza:
  - $MAE^M(h,Te) = \frac{1}{|C|}\sum_{j=1}^{|C|}\frac{1}{Te_j}\sum_{x_i \in Te}|h(x_i) - y_i|$
  - $MAE^\mu(h,Te) = \frac{1}{|Te|}\sum_{x_i \in Te}|h(x_i) - y_i|$

5. Risultati

5.1 Task B

	Categorie	Concetti	Categorie + Concetti
Avg Rec	0.494	0.452	**0.496**

Come si evince dalla tabella sopra riportata, i risultati migliori sono stati ottenuti utilizzando la composizione tra tweet e categorie. L'AvgRec ottenuta si sarebbe piazzata alla posizione 23 su 24 partecipanti se avessimo partecipato al Semeval 2017.

5.2 Task C

	Categorie	Concetti	Categorie + Concetti
MAE^M	4.223	**3.756**	4.192
MAE^μ	0.556	**0.489**	0.551

6. Possibili migliorie

Si propongono due possibili migliorie per la predizione:

1. **Confrontare vari classificatori**: altri classificatori potrebbero fornire risultati migliori
2. **Cambio scelta test set**: il test set potrebbe essere scelto con altri criteri da quello utilizzato
3. **Cross validation**: ripetere l'esperimento varie volte potrebbe migliorare i risultati ottenuti.