DATA MINING PROJECT

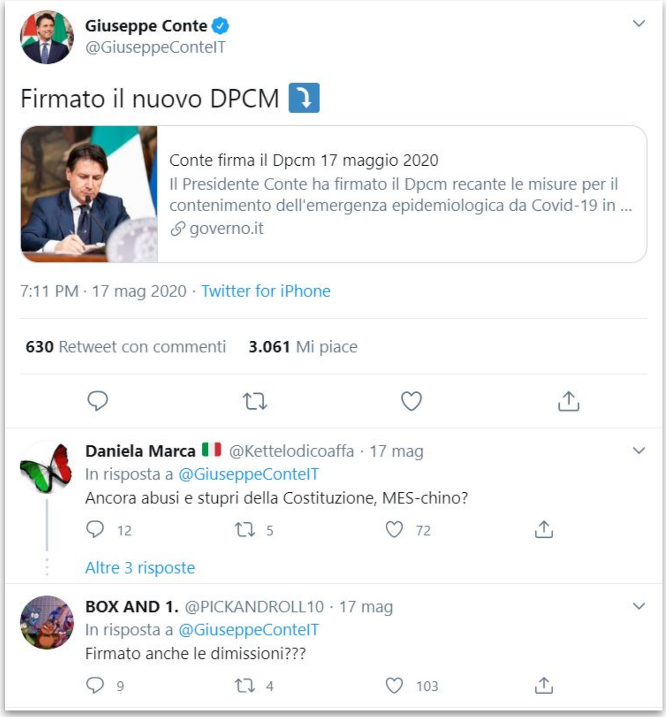
Social Media Listening based on Sentiment Analysis Models for Compliance Evaluation

Luca Barsellotti

INTRODUCTION

The **regulatory compliance** in a community can be strongly correlated with the **percentage of contrast towards the political leader.** The aim of this project is to create a system that allows the monitoring of this percentage of contrast by analysing the **replies on Twitter** to specific tweets posted by political leaders.

TWEETS SCRAPING



Giuseppe Conte @GiuseppeConteIT

Firmato il nuovo DPCM ⤵️

Conte firma il Dpcm 17 maggio 2020
Il Presidente Conte ha firmato il Dpcm recante le misure per il contenimento dell'emergenza epidemiologica da Covid-19 in ...
🔗 governo.it

7:11 PM · 17 mag 2020 · Twitter for iPhone

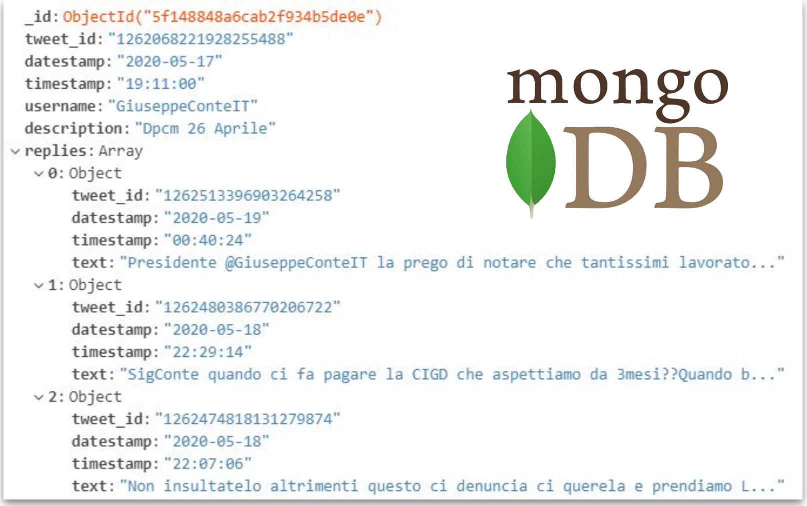
630 Retweet con commenti **3.061** Mi piace

Daniela Marca 🇮🇹 @Kettelodicoaffa · 17 mag
In risposta a @GiuseppeConteIT
Ancora abusi e stupri della Costituzione, MES-chino?
12 5 72

Altre 3 risposte

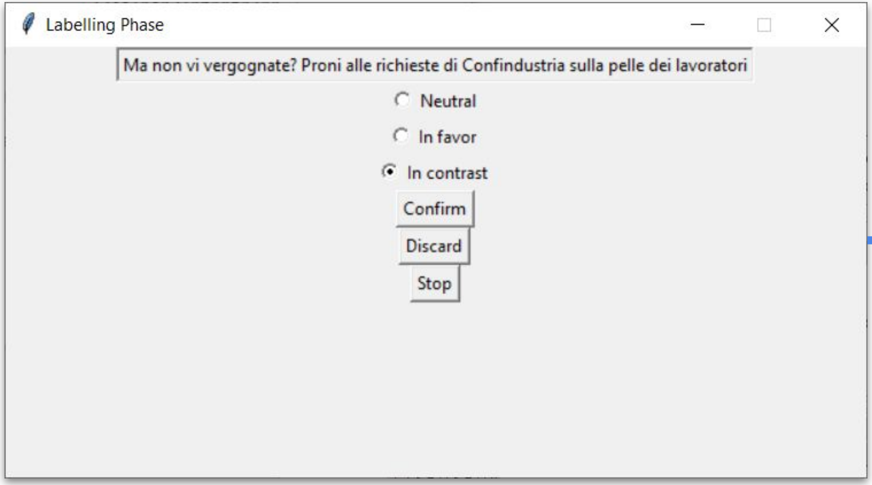
BOX AND 1. @PICKANDROLL10 · 17 mag
In risposta a @GiuseppeConteIT
Firmato anche le dimissioni???
9 4 103

Twint



```
_id: ObjectId("5f148848a6cab2f934b5de0e")
tweet_id: "1262068221928255488"
datestamp: "2020-05-17"
timestamp: "19:11:00"
username: "GiuseppeConteIT"
description: "Dpcm 26 Aprile"
⌄ replies: Array
  ⌄ 0: Object
      tweet_id: "1262513396903264258"
      datestamp: "2020-05-19"
      timestamp: "00:40:24"
      text: "Presidente @GiuseppeConteIT la prego di notare che tantissimi lavorato..."
  ⌄ 1: Object
      tweet_id: "1262480386770206722"
      datestamp: "2020-05-18"
      timestamp: "22:29:14"
      text: "SigConte quando ci fa pagare la CIGD che aspettiamo da 3mesi??Quando b..."
  ⌄ 2: Object
      tweet_id: "1262474818131279874"
      datestamp: "2020-05-18"
      timestamp: "22:07:06"
      text: "Non insultatelo altrimenti questo ci denuncia ci querela e prendiamo L..."
```

TRAINING SET CONSTRUCTION

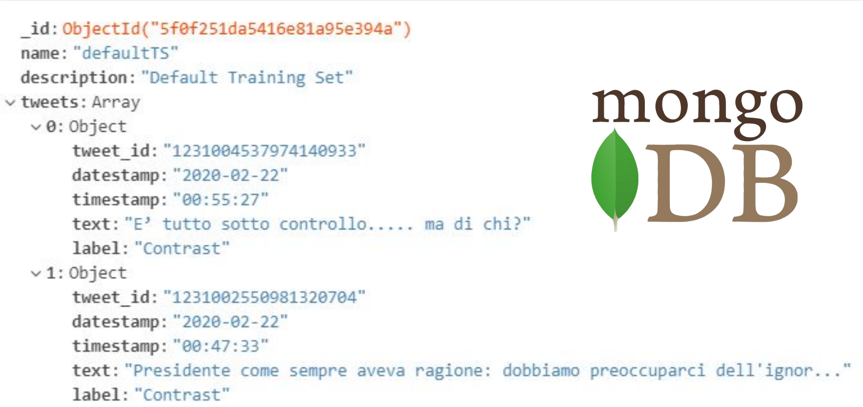


Ma non vi vergognate? Proni alle richieste di Confindustria sulla pelle dei lavoratori

○ Neutral

○ In favor

● In contrast



```
_id: ObjectId("5f0f251da5416e81a95e394a")
name: "defaultTS"
description: "Default Training Set"
⌄ tweets: Array
  ⌄ 0: Object
      tweet_id: "1231004537974140933"
      datestamp: "2020-02-22"
      timestamp: "00:55:27"
      text: "E’ tutto sotto controllo..... ma di chi?"
      label: "Contrast"
  ⌄ 1: Object
      tweet_id: "1231002550981320704"
      datestamp: "2020-02-22"
      timestamp: "00:47:33"
      text: "Presidente come sempre aveva ragione: dobbiamo preoccuparci dell'ignor..."
      label: "Contrast"
```

February 21st

March 14th

Contrast: 320

Neutral: 238

Favour: 253

DATA PREPROCESSING

1. **Transform emojis**: 👏⟶:clapping_hands_sign:⟶clappinghandssign
2. **Lowercase**: “Conte” ⟶ “conte”
3. **Expand abbreviations**: “nn” ⟶ “non”
4. **Cleaning**: “@giuseppeconteit, #dimettiti”⟶”dimettiti”
5. **Stopwords removing**: “a voi sta a cuore la vostra sedia”⟶”cuore sedia”
6. **Stemming**: “cuore sedia”⟶”cuor sed”
7. **Label replacing**:

    “Contrast”⟶-1

    “Neutral”⟶0

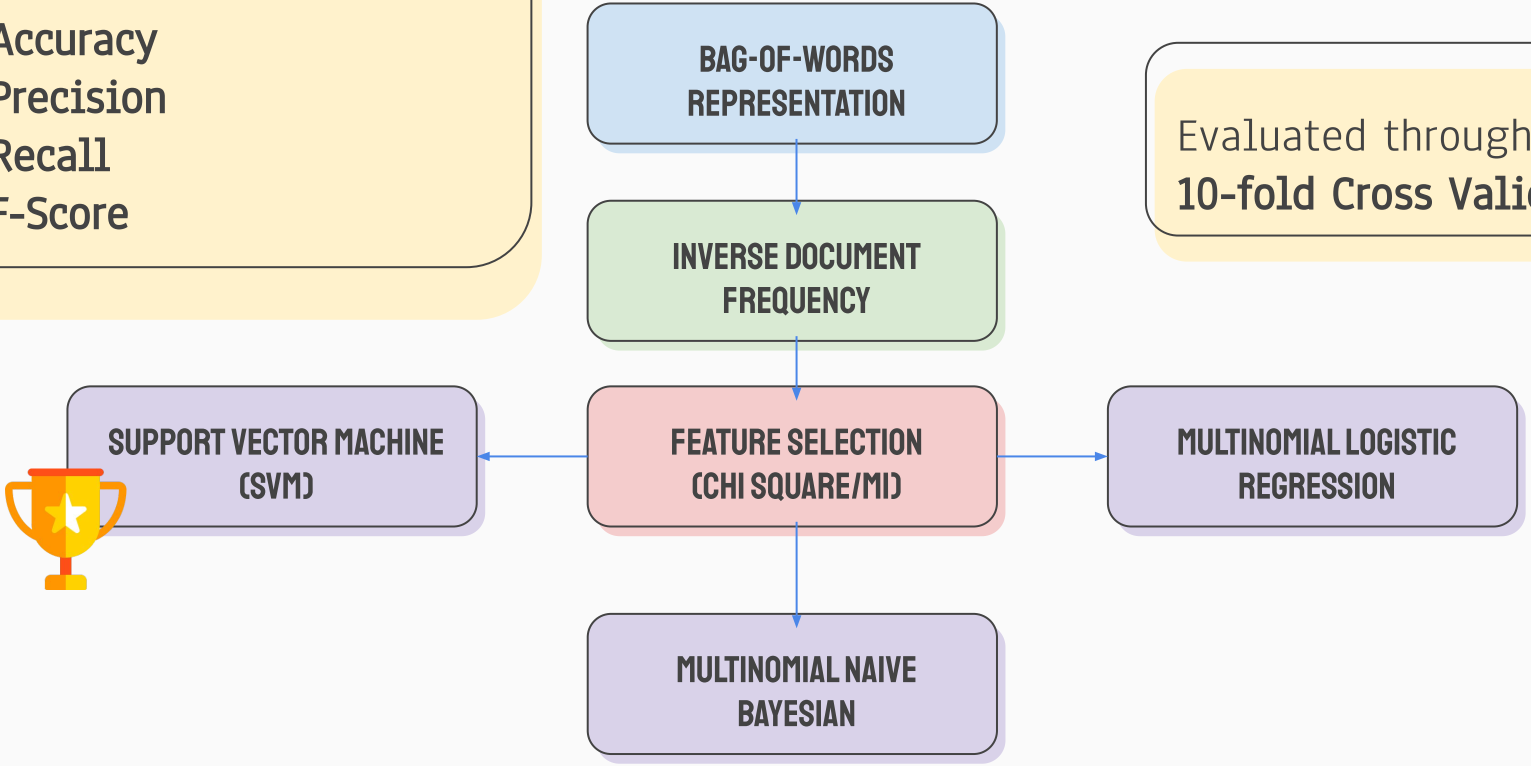
    “Favour”⟶1

MODEL SELECTION: PIPELINE

Adopted metrics:

1. Accuracy
2. Precision
3. Recall
4. F-Score

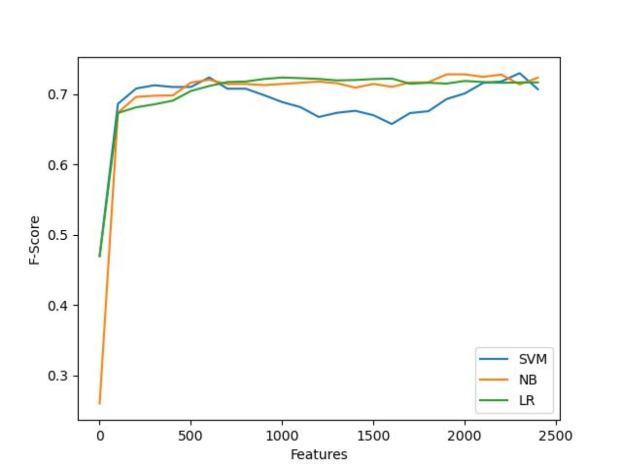
Evaluated through a **10-fold Cross Validation**



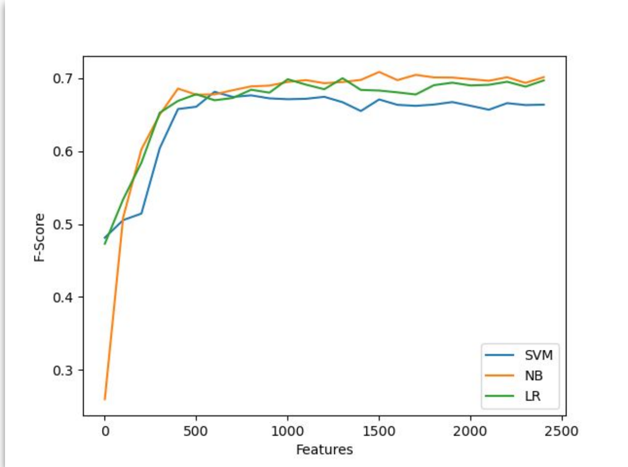
BAG-OF-WORDS REPRESENTATION
INVERSE DOCUMENT FREQUENCY
FEATURE SELECTION (CHI SQUARE/MI)
SUPPORT VECTOR MACHINE (SVM)
MULTINOMIAL LOGISTIC REGRESSION
MULTINOMIAL NAIVE BAYESIAN

Hop = 100

MODEL SELECTION: CHI SQUARE OR MI?



VS



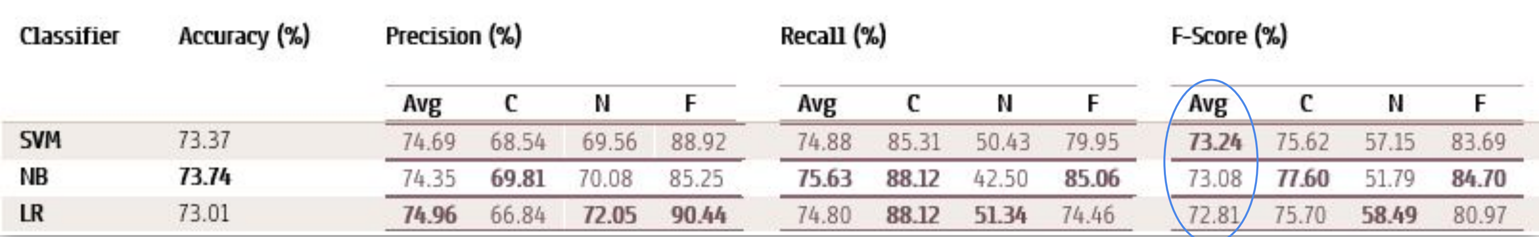
Best result for **Chi Square**:

1. Classifier: **SVM**
2. Features: **2301**
3. F-Score: **72.95%**

Best result for **MI**:

1. Classifier: **NB**
2. Features: **1501**
3. F-Score: **70.88%**

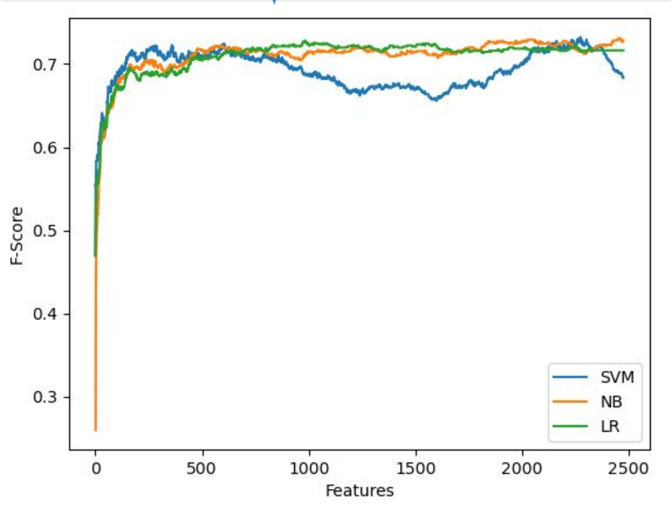
MODEL SELECTION: BEST MODEL



Classifier	Accuracy (%)	Precision (%)				Recall (%)				F-Score (%)			
		Avg	C	N	F	Avg	C	N	F	Avg	C	N	F
SVM	73.37	74.69	68.54	69.56	88.92	74.88	85.31	50.43	79.95	**73.24**	75.62	57.15	83.69
NB	**73.74**	74.35	**69.81**	70.08	85.25	**75.63**	**88.12**	42.50	**85.06**	73.08	**77.60**	51.79	**84.70**
LR	73.01	**74.96**	66.84	**72.05**	**90.44**	74.80	**88.12**	**51.34**	74.46	72.81	75.70	**58.49**	80.97

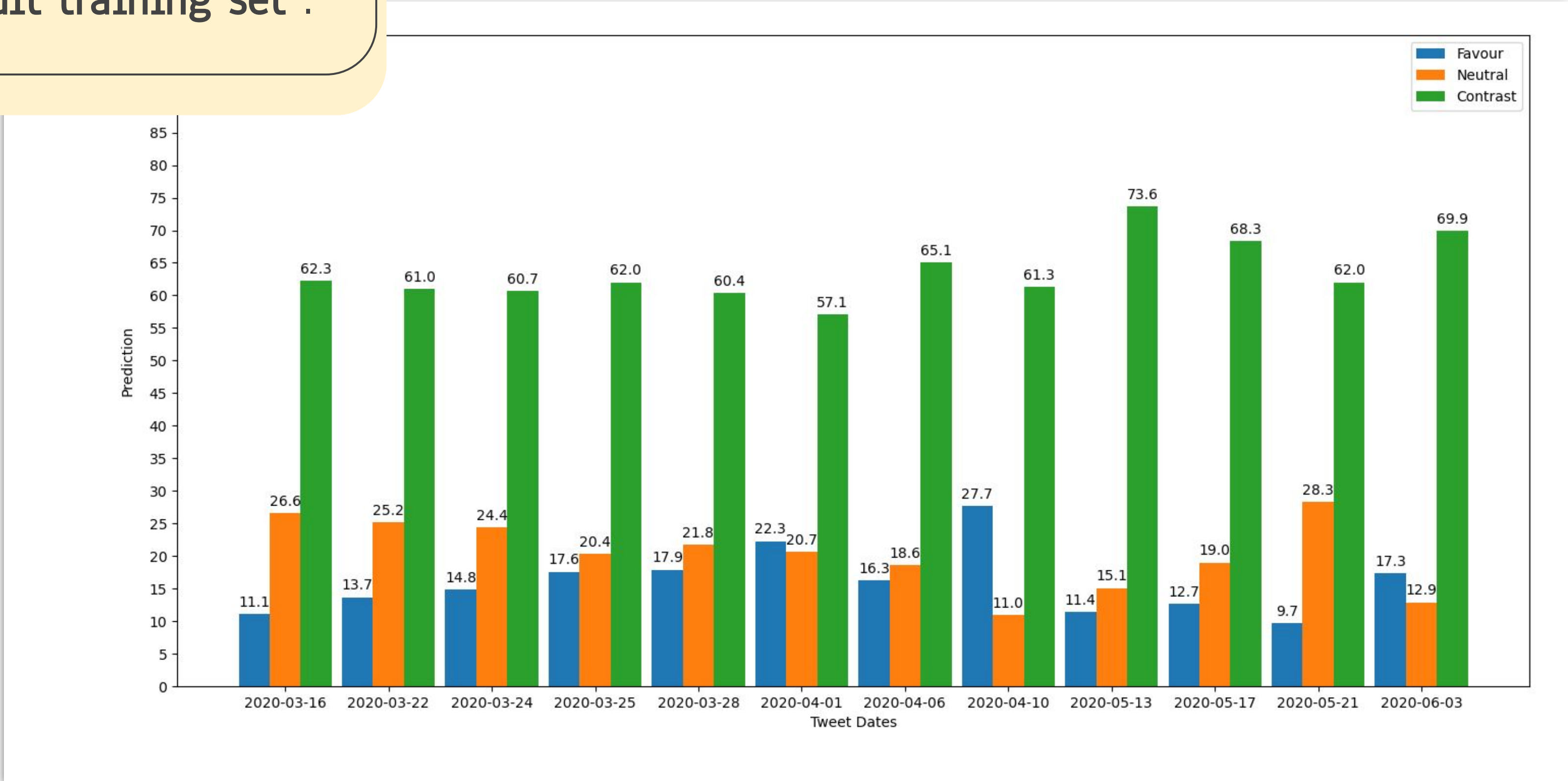
Optimal number of features:

1. SVM→2274
2. NB→2448
3. LR→983



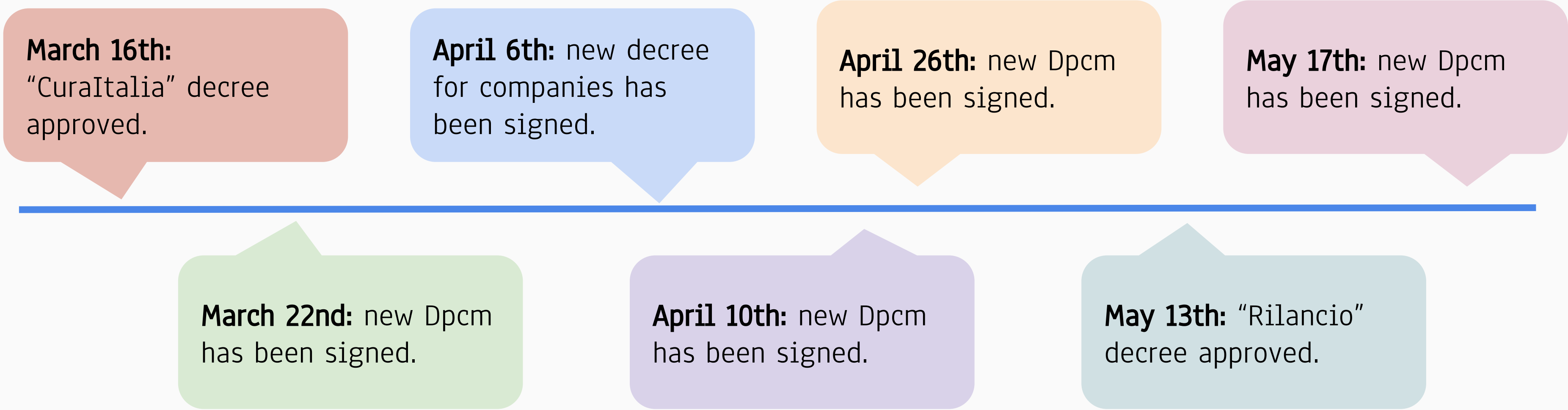
PREDICTION WITH THE DEFAULT TRAINING SET

Prediction through a **SVM model** with **2274** features built on the “**default training set**”.



INCREMENTAL ANALYSIS: CHOSEN EVENTS

Seven events, with the correspondent Giuseppe Conte's tweets, have been chosen to perform an **incremental analysis**.



March 16th: "CuraItalia" decree approved.

March 22nd: new Dpcm has been signed.

April 6th: new decree for companies has been signed.

April 10th: new Dpcm has been signed.

April 26th: new Dpcm has been signed.

May 13th: "Rilancio" decree approved.

May 17th: new Dpcm has been signed.

INCREMENTAL ANALYSIS: LABELS DISTRIBUTION

Event 1:
Contrast: 17
Neutral: 10
Favour: 39

Event 2:
Contrast: 22
Neutral: 11
Favour: 33

Event 3:
Contrast: 12
Neutral: 11
Favour: 46

Event 4:
Contrast: 9
Neutral: 35
Favour: 34

Event 5:
Contrast: 25
Neutral: 8
Favour: 30

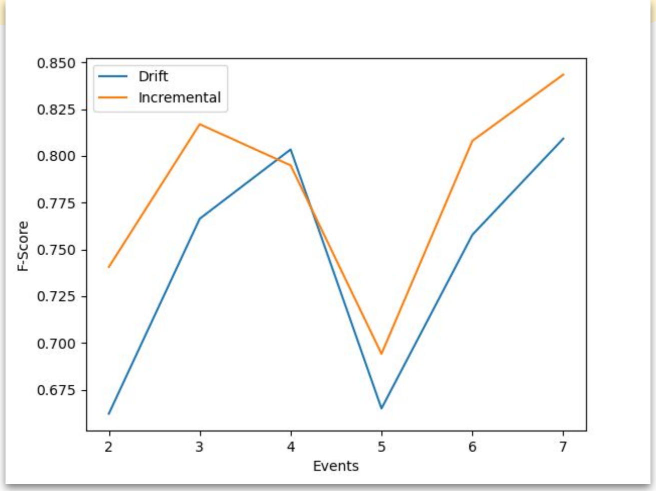
Event 6:
Contrast: 14
Neutral: 9
Favour: 46

Event 7:
Contrast: 16
Neutral: 11
Favour: 39

INCREMENTAL ANALYSIS: F-SCORE

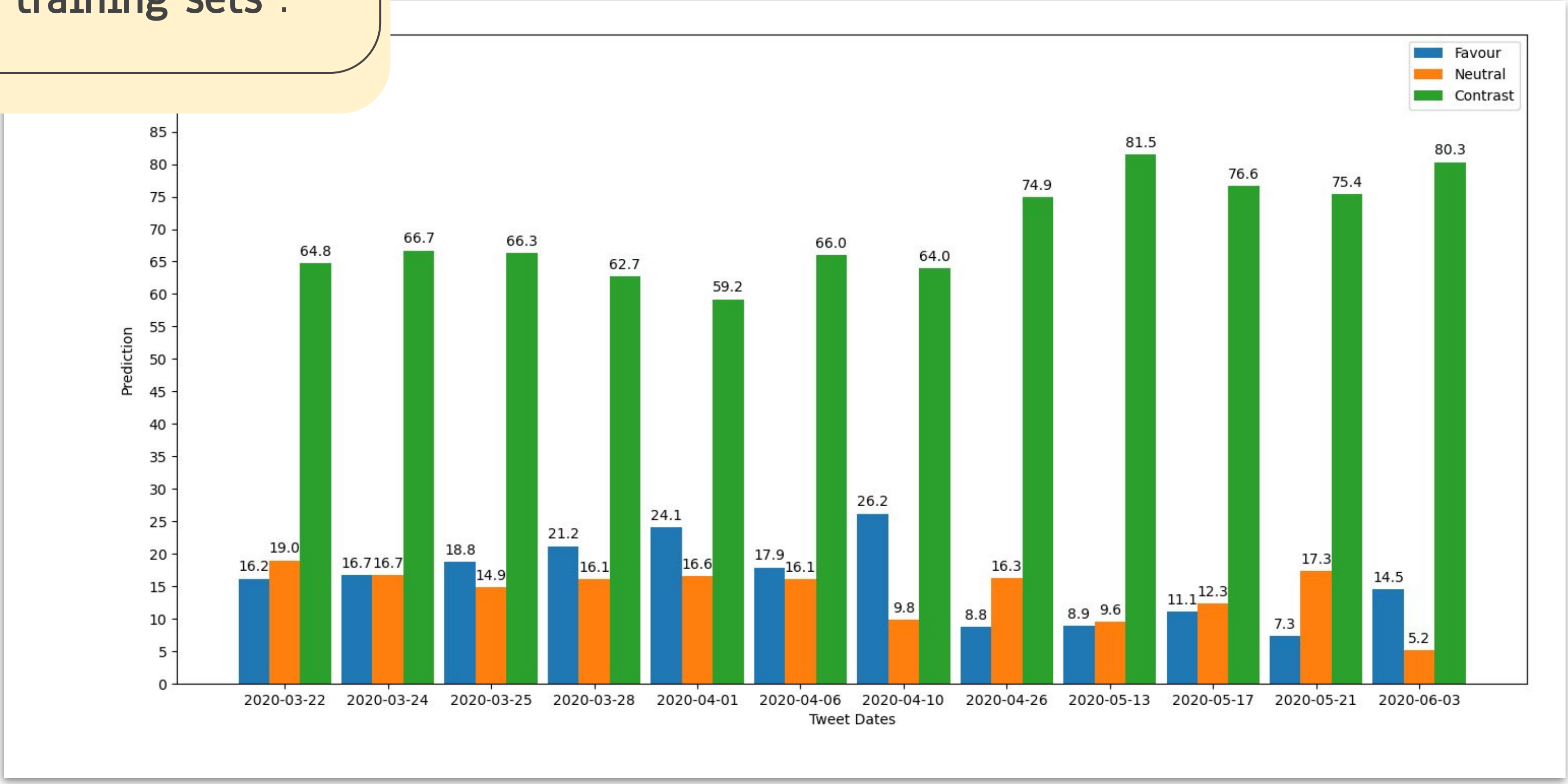
The adopted metrics have been calculated on each “**event test set**” at step *i* using models derived from “**default training set**” and “**incremental training set**” at step *i-1*. For each “**incremental model**”, the best number of features has been found using the Pipeline.

The chart shows the comparison between the two models based on the **average weighted F-Score**.



PREDICTION WITH THE INCREMENTAL TRAINING SET

Prediction through the several optimal models built on the “**incremental training sets**”.



THANK YOU!