# Identification of structural problems using tweets

Data Mining 2019/2020

Matteo Suffredini

# Context

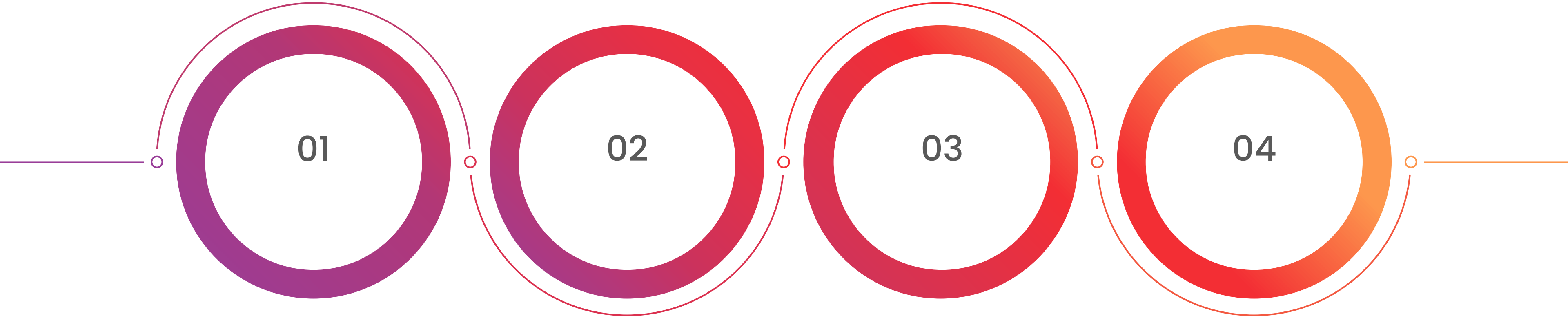
## Problem

Identify structural problem so that they can be solved promptly

## Solution

Use tweets by selecting those that relate to structural problems

# ROADMAP



01

## Obtain tweets

download tweets and clean them

02

## Text elaboration

allows to obtain a features rappresentation

03

## Train classifier

several experiments to obtain the best

04

## Text classification

classifies new tweets using the classifier

# Obtain tweets

## Tool

GetOldTweet
Library

## Search

Position
(LAT, LON, Range)

Keywords

## Cleaning

Remove duplicates

Convert uppercase

Remove useless
meta-informations

# Text Elaboration

## Tokenization

<lo>, <scarico>, <in>, <via>, <verdi>, <perde>

## Stop-word filtering

<scarico>,<via>, <verdi>, <perde>

## Stemming

<scaric>,<vi>, <verd>, <perd>

## Stem filtering

Select F relevant words with positive IG (IDF assigned to words)

## Feature rapresentation

$X = [\ Xvi, Xscaric, Xfogn, Xacqu, Xrottur, Xriparazion, \dots\ ]_F$

$X_i = [\ Wvi, 0, Wfogn, 0, 0, 0, \dots\ ]_F$

# Experiment

## Validation

10 Stratified
Cross Fold

2 seed

## Classifier

Decision Tree
SVM
Multinomial NB
K-NN
Adaboost
RandomForest

## Metrics

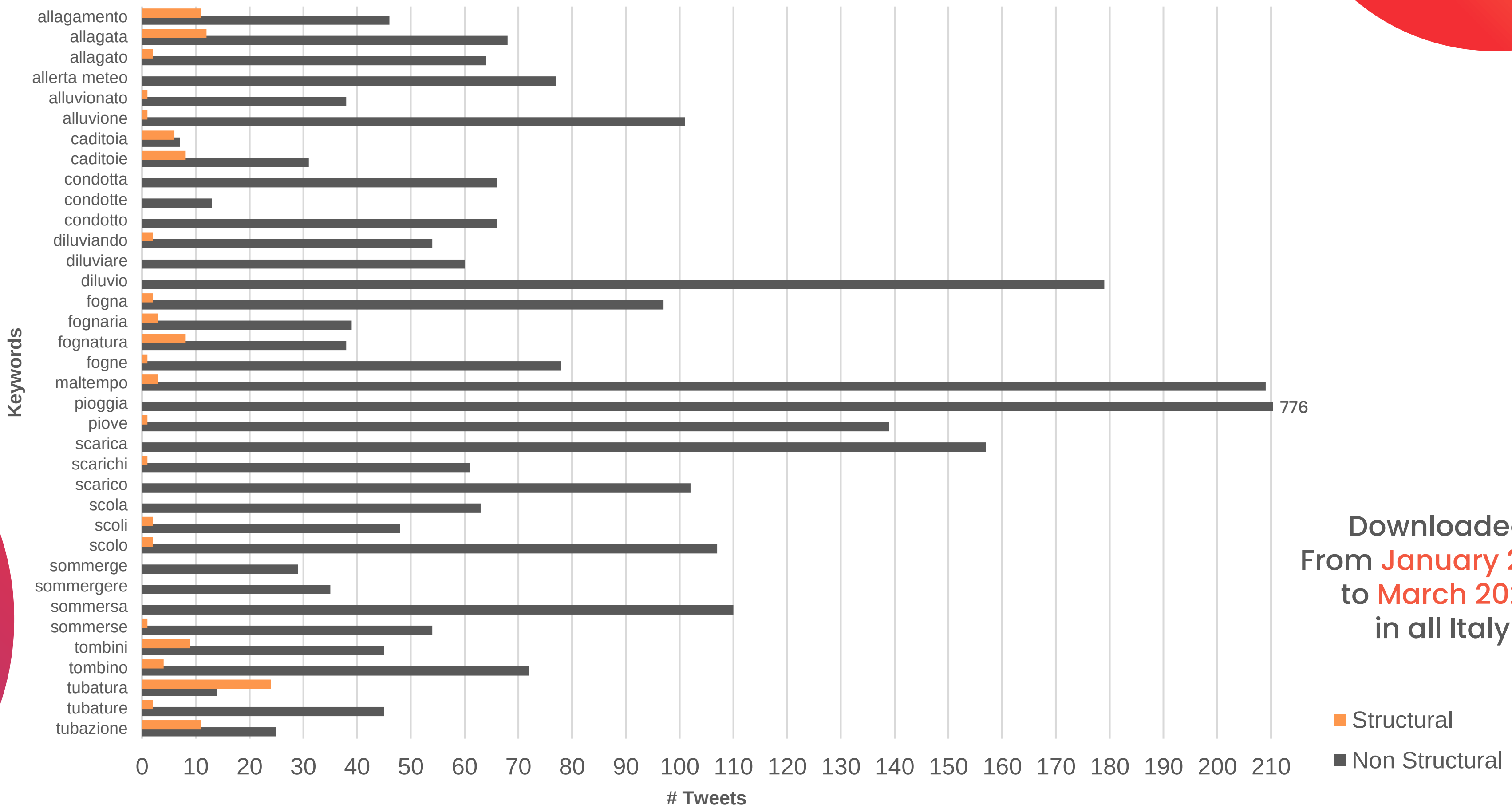
Accuracy
Precision
Recall
F-Score

# Experiment

Deal with unbalanced dataset

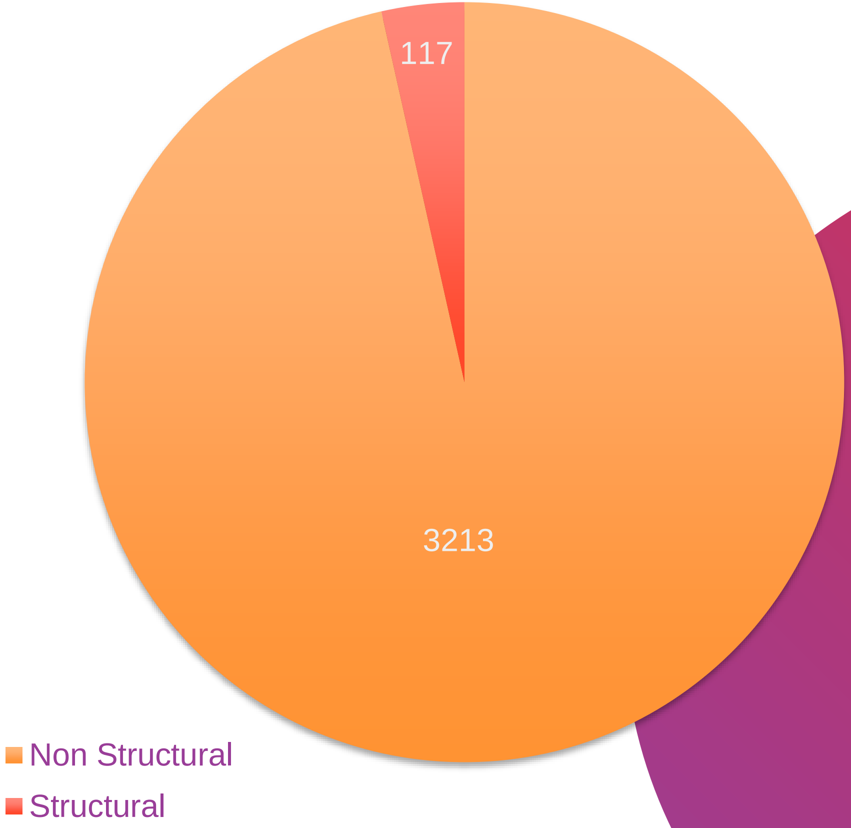
01

# Tweets distribution per keyword



Downloaded
From January 2000
to March 2020
in all Italy

# Dataset

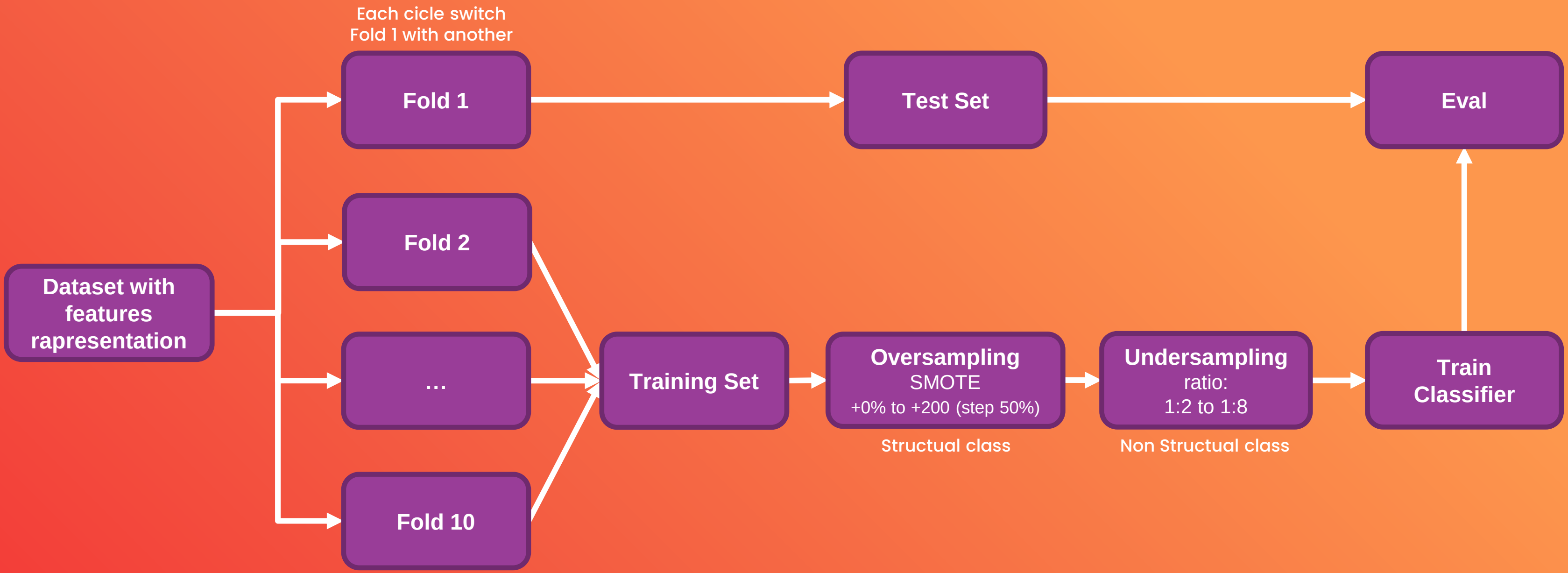


**Deal with unbalanced dataset**

Oversampling
**Structural** class

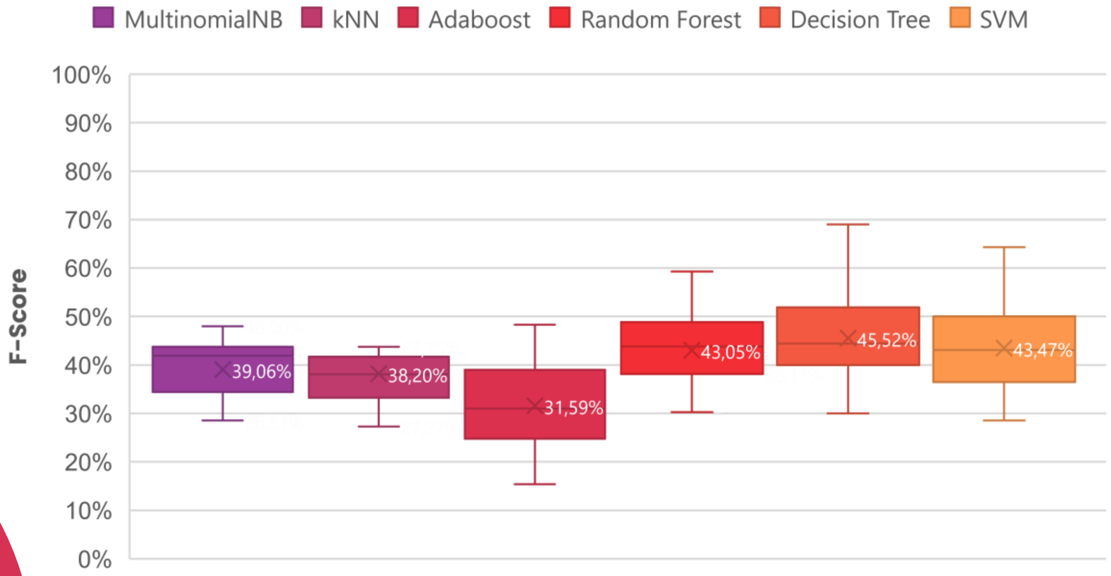
Undersampling
**NonStructural** class

# Evaluation example



The oversampling % and undersampling ratio have been varied together with the specific parameters of each classifier in order to obtain the best average f-score

# Experiment 1 results



■ **MultinomialNB**
Over: 150%  Under: 1:4

■ **K-NN**
Over: 50%  Under: 1:4  k: 1

■ **Adaboost**
Over: 0%  Under: 1:8

■ **Random Forest**
Over: 50%  Under: 1:4

■ **Decision Tree**
Over: 0%  Under: 1:5

■ **SVM**
Over: 50%  Under: 1:5  $W_{struct}$: 0,5

Selected Classifier

## Decision Tree

| | |
|---|---|
| **Accuracy:** | 95,74 ± 0,47 |
| **F-Score:** | 45,52 ± 4,93 |
| **Precision:** | 42,30 ± 4,71 |
| **Recall:** | 51,36 ± 7,12 |

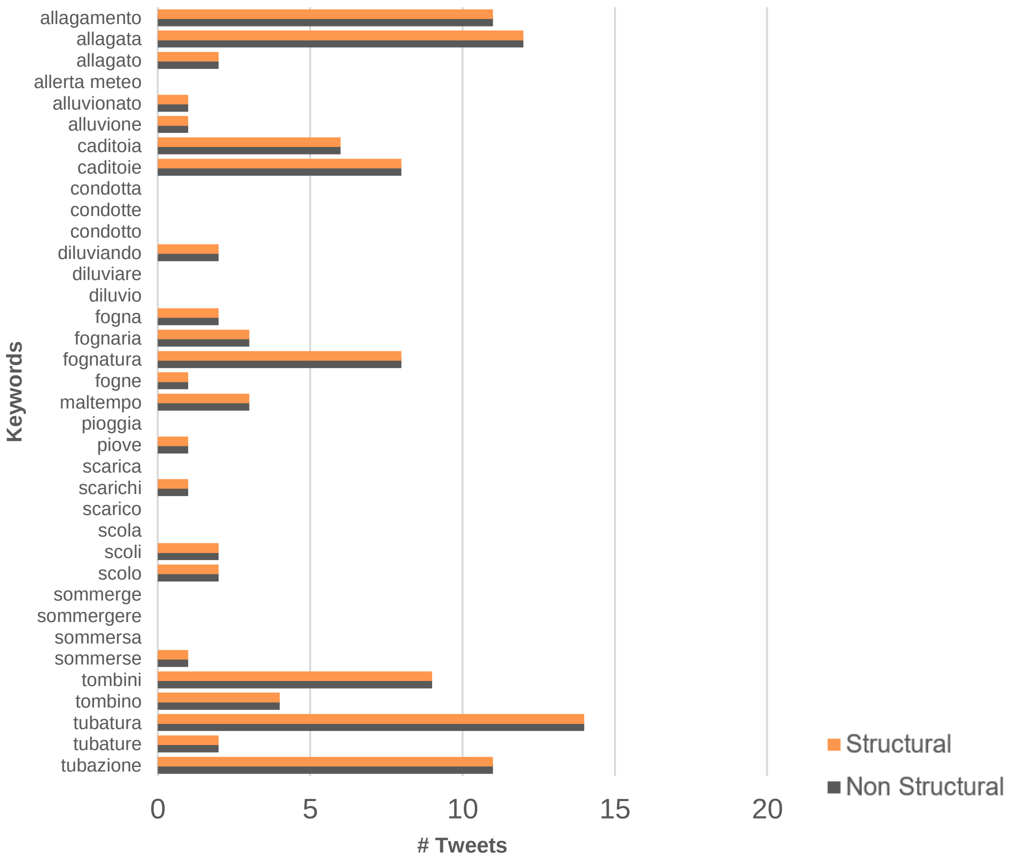
| ●●●●●● | Structural | Non Structural | Tot |
|---|---|---|---|
| Structural | 120 | 114 | 234 |
| Non Structural | 170 | 6256 | 6426 |
| Tot | 290 | 6370 | **6660** |

# Experiment

Resample per word

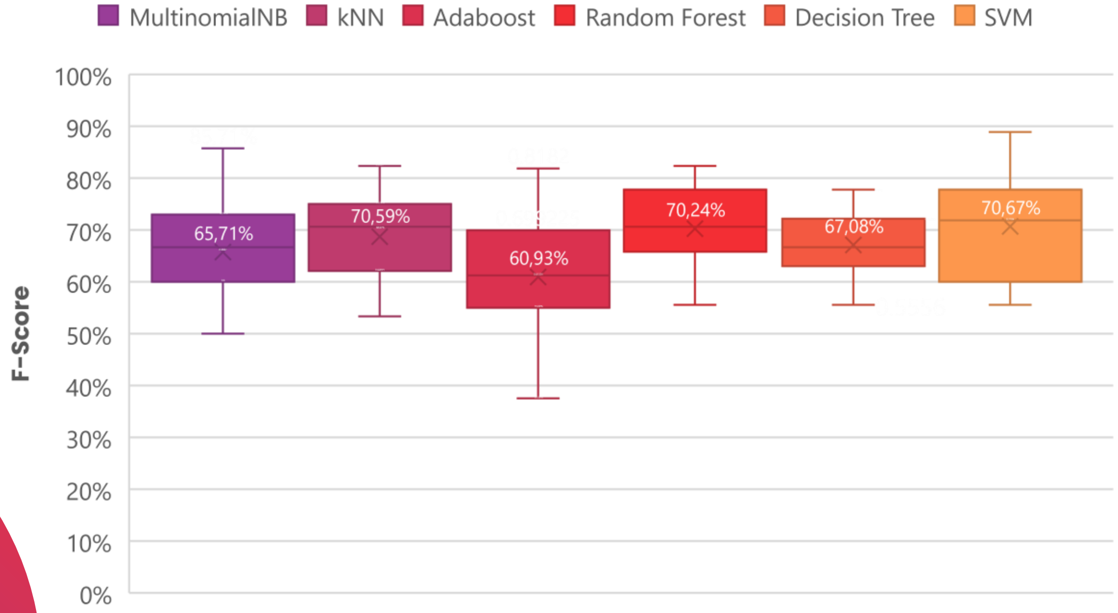
02

# Tweets distribution per keyword



214 Tweets Selected from the original dataset

# Experiment 2 results



🟥 **K-NN**
k: 1

🟧 **SVM**
$W_{struct}$: 0,6

Selected Classifier

**SVM**

**Accuracy:** 72,47 ± 4,82

**F-Score:** 70,67 ± 4,25

**Precision:** 80,15 ± 7,09

**Recall:** 65,50 ± 5,42

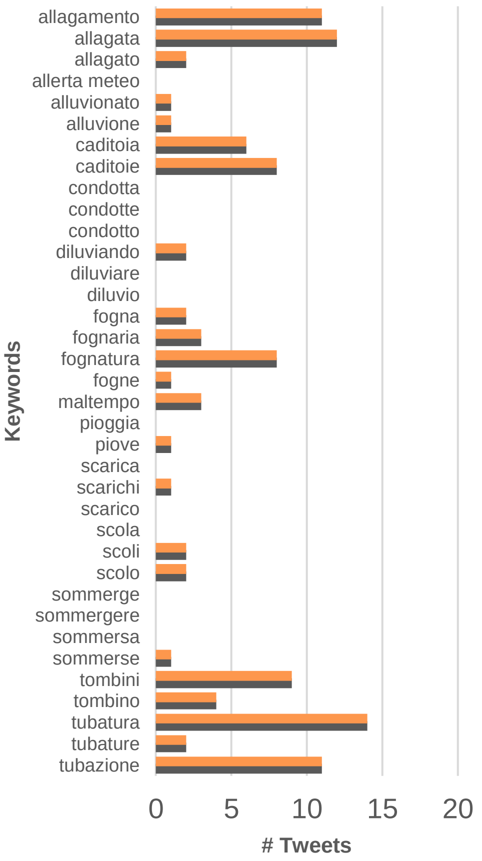
| ●●●●●● | Structural | Non Structural | Tot |
|---|---|---|---|
| Structural | 140 | 74 | 214 |
| Non Structural | 44 | 170 | 214 |
| Tot | 184 | 244 | **428** |

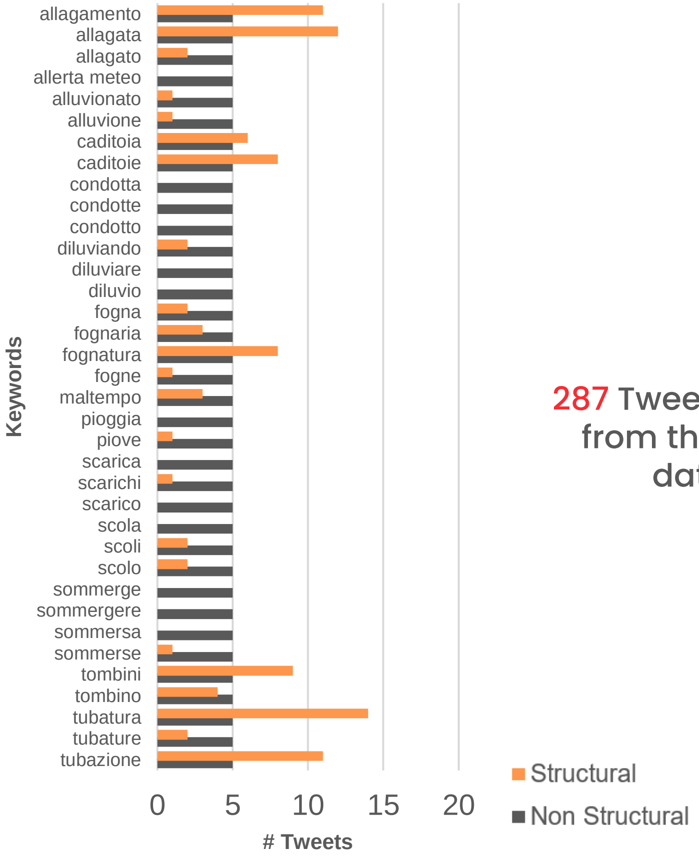
03

# Experiment

Resample per word with constant Non Structural tweets

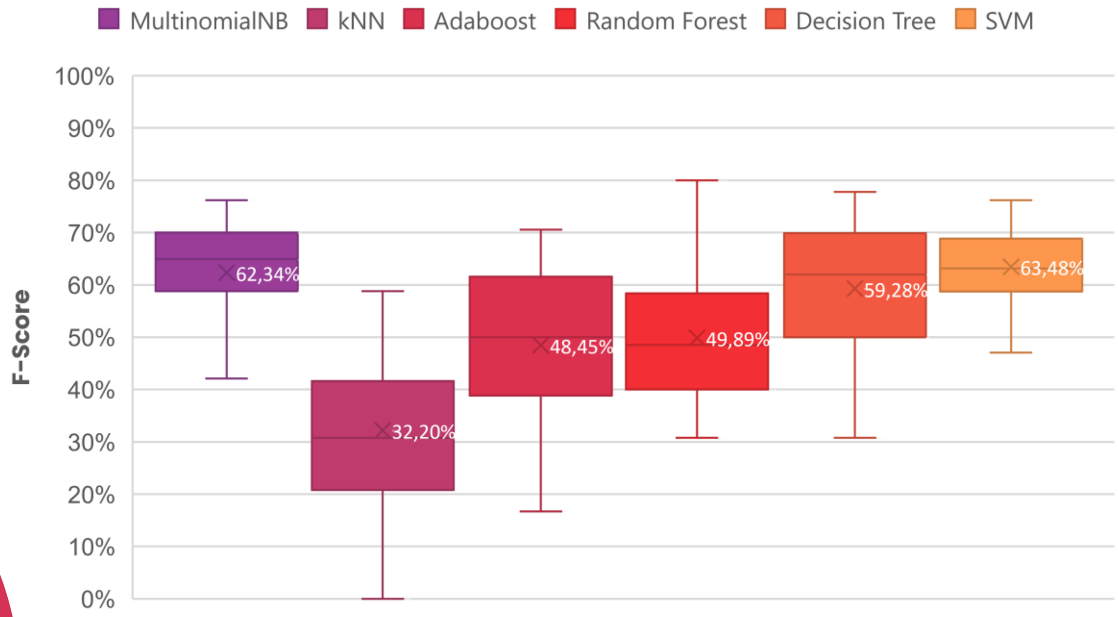
# Tweets distribution per keyword





287 Tweets Selected from the original dataset

# Experiment 3 results



🟥 **K-NN**
k: 1

🟧 **SVM**
W<sub>struct</sub>: 0,8

Selected Classifier

**SVM**

| | |
|---|---|
| **Accuracy:** | 73,36 ± 2,63 |
| **F-Score:** | 63,48 ± 4,03 |
| **Precision:** | 64,89 ± 3,37 |
| **Recall:** | 63,55 ± 6,15 |

| ●●●●●● | Structural | Non Structural | Tot |
|---|---|---|---|
| Structural | 136 | 78 | 214 |
| Non Structural | 75 | 285 | 360 |
| Tot | 211 | 363 | **574** |

# Experiment 2 vs 3

**Test classifier build with all selected tweets with excluded tweets from original dataset**

## SVM

Experiment 2

|  | S | NS |  |
|---|---|---|---|
| **S** | 7 | 3 | 10 |
| **NS** | 667 | 2439 | 3106 |
|  | 674 | 2442 | **3116** |

**Accuracy:** 78,50%
**Recall NS:** 78,53%
**Recall S:** 70,00%

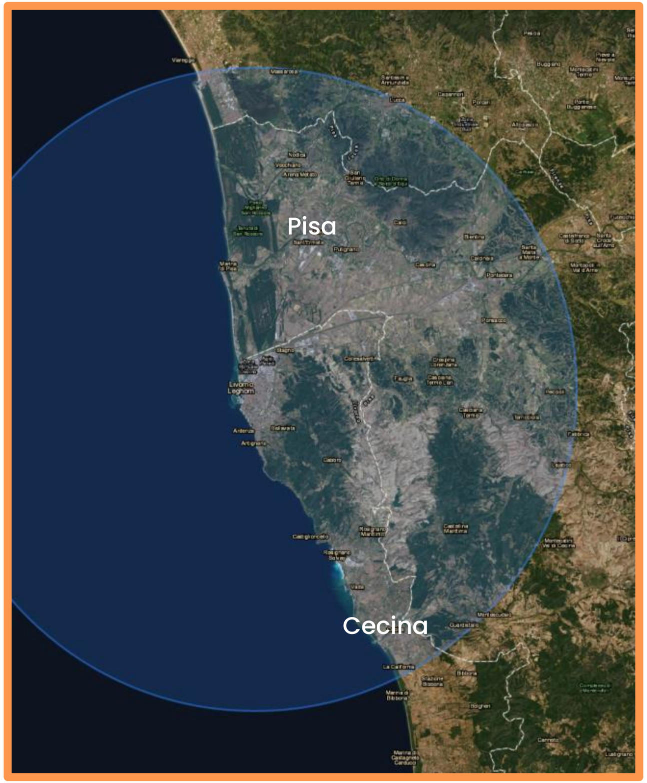
## SVM

Experiment 3

|  | S | NS |  |
|---|---|---|---|
| **S** | 8 | 2 | 10 |
| **NS** | 307 | 2726 | 3033 |
|  | 315 | 2728 | **3043** |

**Accuracy:** 89,85%
**Recall NS:** 89,88%
**Recall S:** 80,00%

# Comparison with heavy rainfall events



Center: 43,5436 - 10,317
Radius: 37 km ≈ 23 mi

| Event | | | Classific. Results | | Real | | TOT | |
|---|---|---|---|---|---|---|---|---|
| Data e Ora Inizio | Data e Ora Fine | Luogo | S | NS | S | NS | Keys | No Key |
| 04/04/2015 12:45 | 04/04/2015 19:00 | CECINA | 0 | 5 | 0 | 5 | 5 | 146 |
| 8/24/15 3:00 | 8/24/15 7:00 | PISA | 1 | 1 | 1 | 1 | 2 | 93 |
| 10/27/15 9:00 | 10/28/15 15:30 | PISA + CECINA | 0 | 0 | 0 | 0 | 0 | 145 |
| 08/05/2016 09:00 | 08/05/2016 11:00 | PISA | 0 | 1 | 0 | 1 | 1 | 116 |
| 9/16/16 7:45 | 9/16/16 13:30 | CECINA | 0 | 4 | 0 | 4 | 4 | 113 |
| 10/14/16 13:00 | 10/14/16 21:00 | PISA + CECINA | 0 | 2 | 0 | 2 | 2 | 94 |
| 09/09/2017 20:00 | 09/10/2017 09:00 | PISA + CECINA | 0 | 2 | 0 | 2 | 2 | 112 |
| 9/18/17 12:30 | 9/19/17 9:30 | CECINA | 0 | 0 | 0 | 0 | 0 | 101 |
| 12/10/2017 09:00 | 12/11/2017 09:00 | PISA | 0 | 1 | 0 | 1 | 1 | 96 |
| 10/28/18 9:00 | 10/29/18 9:00 | PISA + CECINA | 0 | 1 | 0 | 1 | 1 | 94 |
| 2/17/18 14:45 | 2/18/18 7:30 | PISA + CECINA | 0 | 0 | 0 | 0 | 0 | 94 |
| 10/24/19 10:30 | 10/24/19 23:45 | PISA | 0 | 4 | 0 | 4 | 4 | 175 |
| 1/27/20 7:15 | 1/27/20 13:45 | PISA + CECINA | 0 | 1 | 0 | 1 | 1 | 212 |
| 10/29/19 9:00 | 10/30/19 9:00 | PISA + CECINA | 0 | 5 | 0 | 5 | 5 | 175 |
| | | TOT | 1 | 27 | 1 | 27 | 28 | 1766 |

Analysing a subset of the 1766 tweets, there do not seem to be any structural tweets not related to the words used for the search

# Tool

Structural Tweet visualization

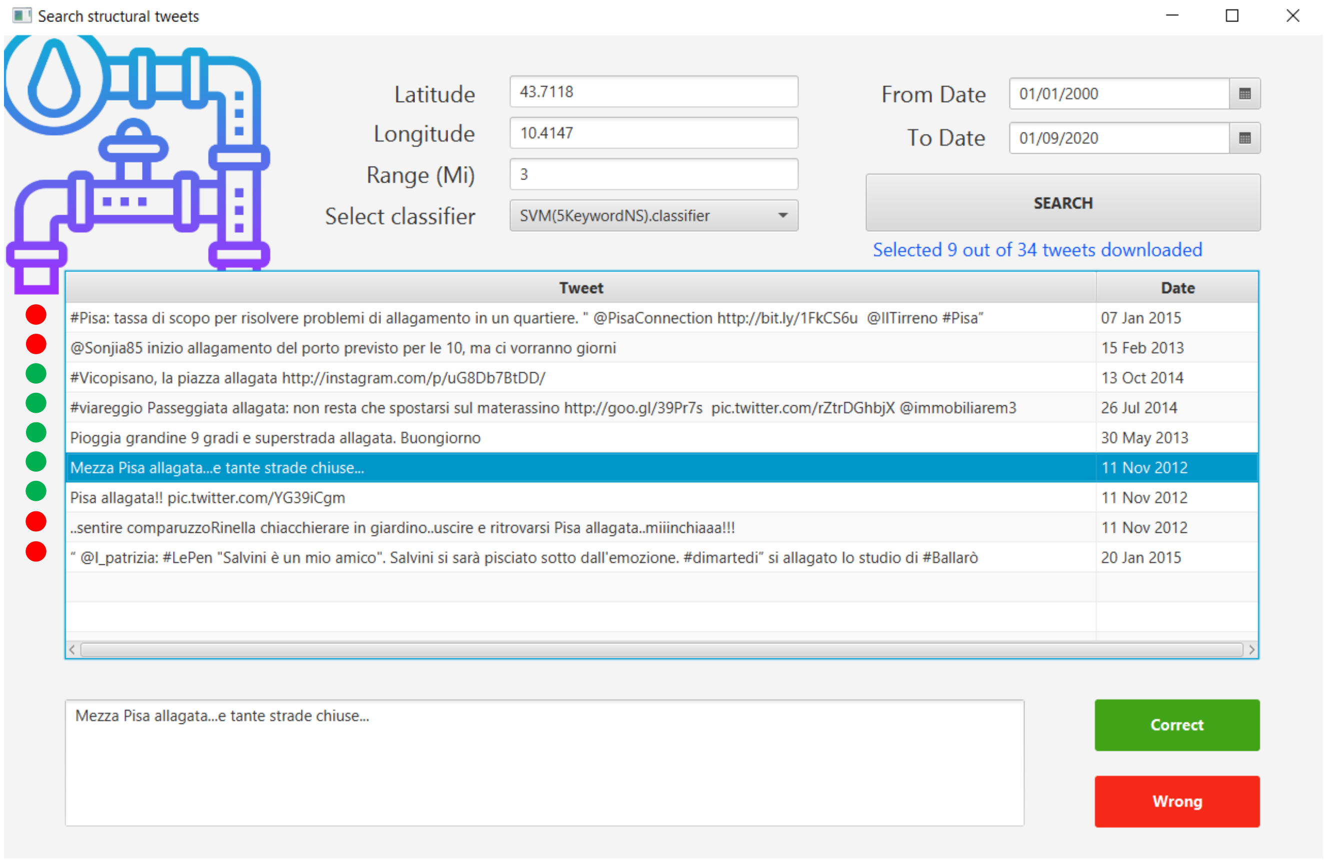
Terme
Ascia
arino
ore
Cascine
Vecchie
Pisa
Sant'Ermete
Putignano
Luicchio
La Vettola
Riglione-
Oratoio
San Piero
a Grado
Na

Radius: 3mi ≈ 5Km

## Output file

| | | |
|---|---|---|
| #Pisa: tassa di scopo … | NonStructural | allagamento |
| @Sonjia85 inizio allagame… | NonStructural | allagamento |
| #Vicopisano, la piazza…. | Structural | allagata |
| #viareggio … | Structural | allagata |
| Pioggia grandine 9 gradi….. | Structural | allagata |
| Mezza Pisa allagata... | Structural | allagata |
| Pisa allagata!! …. | Structural | allagata |
| ..sentire comparuzzoRinella… | NonStructural | allagata |
| “ @l_patrizia: #LePen «… | NonStructural | allagato |

# GUI



Search structural tweets

Latitude: 43.7118

Longitude: 10.4147

Range (Mi): 3

Select classifier: SVM(5KeywordNS).classifier

From Date: 01/01/2000

To Date: 01/09/2020

Selected 9 out of 34 tweets downloaded

| | Tweet | Date |
|---|---|---|
| 🔴 | #Pisa: tassa di scopo per risolvere problemi di allagamento in un quartiere. " @PisaConnection http://bit.ly/1FkCS6u @IlTirreno #Pisa" | 07 Jan 2015 |
| 🔴 | @Sonjia85 inizio allagamento del porto previsto per le 10, ma ci vorranno giorni | 15 Feb 2013 |
| 🟢 | #Vicopisano, la piazza allagata http://instagram.com/p/uG8Db7BtDD/ | 13 Oct 2014 |
| 🟢 | #viareggio Passeggiata allagata: non resta che spostarsi sul materassino http://goo.gl/39Pr7s pic.twitter.com/rZtrDGhbjX @immobiliarem3 | 26 Jul 2014 |
| 🟢 | Pioggia grandine 9 gradi e superstrada allagata. Buongiorno | 30 May 2013 |
| 🟢 | Mezza Pisa allagata...e tante strade chiuse... | 11 Nov 2012 |
| 🟢 | Pisa allagata!! pic.twitter.com/YG39iCgm | 11 Nov 2012 |
| 🔴 | ..sentire comparuzzoRinella chiacchierare in giardino..uscire e ritrovarsi Pisa allagata..miiinchiaaa!!! | 11 Nov 2012 |
| 🔴 | “ @l_patrizia: #LePen "Salvini è un mio amico". Salvini si sarà pisciato sotto dall'emozione. #dimartedi” si allagato lo studio di #Ballarò | 20 Jan 2015 |

Mezza Pisa allagata...e tante strade chiuse...

# Questions?

Thank you for listening