Spotting SARS-COV-2 Misinformation in Italy

Piero Romare

Abstract

Misinformation is not a new phenomenon but the popularity and ubiquity of social media speed up that the amount and velocity at which information is produced and spreads greatly outpaces our ability to evaluate whether it is correct and unbiased. This is especially important in healthcare, where misinformation can influence attitudes and health behaviours that can lead to harm.

In this project, we explore the diffusion of Low Credibility Media on Twitter, during the month of March. We collected over 190.000 Italian tweets relating Covid-19 pandemic. We quantify and analyze tweets which share news from High Credibility Media (HCM) and from Low Credibility Media (LCM). Lastly, we perform classification task between the two classes using different Machine Learning models.

1. Introduction

With social media, people are no longer just passive readers of news media but are also involved in its production and sharing. Web platforms today can diffuse a lot of information in a restricted time and it is difficult filtered the advice based on truth or fake. The safety, usability and reliability of some platforms are compromised by the prevalence of online antisocial behavior that can shape the others opinion [1]. While social media has led to a range of advantages by allowing access to different views, it has also made it easier for misinformation to spread and persist [2]. Vosoughi et al.[3] showed that false news spreads faster and further than true news. Fortunately, it seems that users in the social media tend to prefer real information over false news, in terms of questions from users [4]. The information process is complex, users can generate information either by providing their observations, by bringing relevant knowledge from external sources, by deriving interpretations. Discussions are in a continuous process and during the start of 2020, the worldwide Social Media debate is more about Covid-19 [5].

Misinformation detection is a major challenge. Covid-19 open some interesting developments come from Google [6] and LaRepubblica [7] which has collaborated with the European Parliament. Evaluating the credibility of information presented online often requires specific expertise and a huge amount of time. If we need to rely on expert appraisal to detect false news, it would be impossible to keep up. This is a risk for health information, where the spread of misinformation can have a detrimental effect on people and their communities. Fake news classification is tested in Facebook posts with Bayes Classifier with 74% of test accuracy [8]. In [9] authors proposed an LSTM neural network model that is emotionally infused to detect false news. A survey is elaborated in [10] where there is a distinguish between content based, context based and both using different machine and deep learning models.

In this project, we classify Low Credibility News shared on Twitter in the month of March [11]. We focus on Italian tweets which attach external articles. LCM is not necessarly a fake news, but the source which publish it is already known as a fake news provider. We evaluate different Machine Learning methods that are capable of automatically labelling the credibility of an article [12].

1.1. Contribution

In this project, we make the following contributions:

- We collect Italian scenario tweets on Covid-19
- We quantify the ratio between high and low credibility information on Twitter
- We evaluate different machine learning classifiers to discriminate HCM and LCM using two features extraction modalities

1.2. Organization

This project paper is organised as follows: Section II introduces the data procedure we followed, the model and the evaluation we choose for our aim; Section III the experiments we performed on these data and the results we obtained, in Section IV we summarise our work and discuss the future possibility.

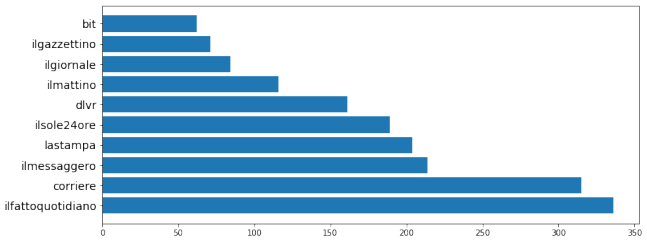


Figure 1. Top 10 Most Common High Media.

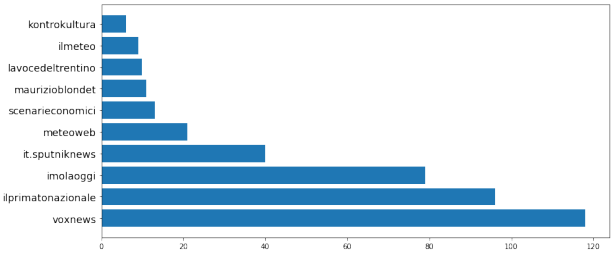


Figure 2. Top 10 Most Common Low Media.

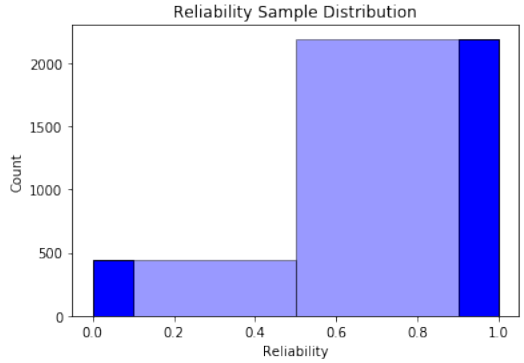


Figure 3. Publish Date.

2. Method

2.1. Data Collection

We randomly collect or hydrate a dataset of 190.000 Italian tweets [11] from March using [13]. We select those tweets which are attached an URL. URLS can be shared as original URL or as shorten URL, in this last case we stretch it to obtain the original domain. We select those tweets which have a domain which is in [14, 15, 16, 17], discarding retweets to avoid the possibility of duplicate articles. We scrape all full body text of articles. Every article has label as reliable article in HCM if its URL domain is in [14] (Fig. 1) and as un-reliable articles in LCM if its URL domain are in [15, 16, 17] (Fig. 2). Here, results that 2192 articles are reliable and 447 articles are un-reliable. In Figure 3 we plot the sample distribution of reliability terms. We show the distribution of publish date of articles scraped (Fig. 6), as you can see there are many missing dates.

2.2. Features Extraction

In order to discriminate HCM and LCM we need to extract useful insights from articles. To do this, we propose two different approaches:

1. *Frequency–Inverse Document Frequency (TF-IDF)*, with an arbitrary stopwords list, defined as:

$$TF_{ij} = \frac{n_{ij}}{|d_j|} \tag{1}$$

$$IDF_i = log \frac{|D|}{\{|d : i \in D|\}} \tag{2}$$

We would highlight the fact that *TF-IDF* is not a proper technique to preprocess features in our investigation. This is because it is based on terms used in the documents and its relative frequency (i.e. if a word is important in a HCM, it could be also in a LCM). It is useful when you might classify different documents topics (e.g. sports vs politics vs science) and, in our case, the documents regarding to a unique topic.

2. Stylometry, where we extract content-based features from title and full body text for every articles (Table 3). We add the *Gulpease Index* [18], an Italian readability index calculates in body text, following the relative scores:

   - < 80 difficult to read for user with elementary degree;
   - < 60 difficult to read for user with middle school degree;
   - < 40 difficult to read for user with high school degree.

   We transform features using *StandardScaler*: $z = \frac{(x-u)}{s}$.

2.3. Evaluations

The dataset is composed with a 1:5 ratio between LCM and HCM. We split training and test sets with a test size of 0.2 and a random seed of 42. The following metrics are used to evaluate models, in particular, we'd like to assign more effort to F1 score and to the Confusion Matrix, because, informally, the most important classification is that a HCM be not classified as LCM (TNR and FPR).

- Recall/Sensitivity/TPR = $\frac{TP}{TP+FN}$;
- Specificity/TNR = $\frac{TN}{TN+FP}$;
- FPR = $\frac{FP}{FP+TN}$;
- F1 $= 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$;

Figure 4. Word Cloud.

- Confusion Matrix;
- Average Precision;
- AUC.

For these evaluations (F1, Average Precision, AUC) we use a micro average: the micro weighs each sample equally.

3. Results

In order to visualize the dataset, we show a Word Cloud based on word frequencies of full body text of the articles of HCM and LCM (Fig. 4) and a word count bar (Fig. 13). We explore different binary machine learning classifier: *Logistic Regression, K Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machine* using Sci-Kit Learn Python Library. We use grid search to find optimal hyperparameters (Table 3-7) with scoring on *precision* and 5-Fold cross validation. We show features importances for Logistic Regression where the number of nouns in the body articles and the number of punctuation in the body articles are the most important (Fig. 7), Decision Tree where the number of words in the title is the most important (Fig. 8) and Random Forest where the importance have a similar rank of Decision Tree features (Fig. 9). The results with stylometry features are provided in Table 2 and the relative confusion matrices (Fig. 10). In Fig. 12-13 we plot comparison of classifiers

Model	TPR	TNR	FPR	F1	AUC	AP
Logistic	0.18	1.00	0.00	0.86	0.59	0.85
KNN	0.42	0.90	0.1	0.82	0.66	0.88
Decision	0.37	0.83	0.17	0.75	0.60	0.86
Random	**0.34**	**0.95**	**0.05**	**0.85**	**0.65**	**0.87**
SVM	0.39	0.90	0.10	0.82	0.65	0.87

Table 1. Models Evaluations based on stylometry features.

Model	TPR	TNR	FPR	F1	AUC	AP
Logistic	0.23	0.97	0.03	0.84	0.60	0.86
KNN	0.18	0.92	0.08	0.80	0.55	0.84
Decision	0.31	0.88	0.12	0.78	0.60	0.86
Random	**0.30**	**0.92**	**0.08**	**0.82**	**0.61**	**0.86**
SVM	0.02	1.00	0.00	0.83	0.51	0.83

Table 2. Models Evaluations based on *TF-IDF* features.

results. As regards instead *TF-IDF* features, the results are provided in Table 8 and the relative confusion matrices (Fig. 11). In Fig. 14-15 we plot comparison of classifiers results.

We consider good results when we got an high as possible sensitivity score which means that the times that a LCM is classified as a LCM, an high specificity score which means that a HCM is classified as HCM, a low FPR which means that a HCM is classified as a LCM. In our case, we choose as mentioned before a micro average for the metrics and we obtain the same value for F1, precision (sensitivity) and recall. Overall, we obtain that *Logistic Regression* and *Random Forest* as best estimator for our task with stylometry features. They both obtain the highest F1 micro score. For what concern the models trained with *TF-IDF* features, the results achieve similar scores to others with stylometry features, we evaluate also here the *Random Forest* as best estimator due the fact that it show a high TPR and a low FPR with a F1 score of 0.82.

4. Conclusion

People have a different point of view (e.g. conspiracy), but till now we do not have an institution or or anything else who is the keeper of truth. It is needed a trade-off between freedom of speech and censuring. The misinformation problems concern both computational and cognition. In this project, we explore a branch of computational linguistic using different machine learning classifier trained with content-based features to classify reliable and un-reliable news articles. With articles related to Covid-19 the best estimator for what concern stylometry features is the *Random Forest Classifier* which show a F1 score of 0.85 and correctly classify a HCM in the 95% of the cases and correctly in $\frac{1}{3}$ of times the LCM. Also *Logistic Regression* give us a F1 score of 0.86 but in the 82% of times misclassify LCM.

4.1. Future Work

As future work, we first plain to extend our dataset by introducing new articles provided in Veneto which we collected from April to June. Furthermore, in the original dataset we collected also context-based features which open some exploration with Machine Learning classifiers. With an unbalanced dataset another investigation could be in the usage of *under sampling* and *over sampling* [19].

This is a part of a larger project in which we've built different directed graphs, inspired by [20, 21], its relative metrics, to quantify the potential impact [22] of spreading LCM. Thus, possible directions: *EXP* and *SIR* epidemiological models to estimate *Echo Chambers* [23], Politics Profiling of user based on which news they share [24], it could be also use for weighting directed graphs and OSINT Analysis.

References

[1] S. Kumar and N. Shah, "False information on web and social media: A survey," 2018.

[2] X. Qiu, D. Oliveira, A. Shirazi, A. Flammini, and F. Menczer, "Limited individual attention and online virality of low-quality information," *Nature Human Behaviour*, vol. 1, p. 0132, 06 2017.

[3] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, pp. 1146–1151, 03 2018.

[4] M. Mendoza, B. Poblete, and C. Castillo, "Twitter under crisis: Can we trust what we rt?," p. 71–79, 2010.

[5] E. Brugnoli, A. L. Schmidt, E. Grassucci, A. Scala, W. Quattrociocchi, and F. Zollo, "The public debate on social media," *Data Science Task Force about online disinformation powered by AGCOM - Servizi Economico-Statistici*, 2020.

[6] "https://toolbox.google.com/factcheck/explorer."

[7] "https://www.repubblica.it/robinson/2020/06/22/news/sul_sito_di_repubblica_nasce_true_per_combattere_le_fake_news-259920578/."

[8] M. Granik and V. Mesyura, "Fake news detection using naive bayes classifier," in *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, pp. 900–903, 2017.

[9] B. Ghanem, P. Rosso, and F. Rangel Pardo, "An emotional analysis of false information in social media and news articles," *ACM Transactions on Internet Technology*, vol. 20, pp. 1–18, 04 2020.

[10] F. Pierri and S. Ceri, "False news on social media: A data-driven survey," *SIGMOD Rec.*, vol. 48, p. 18–27, Dec. 2019.

[11] E. Chen, K. Lerman, and E. Ferrara, "Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set," *JMIR Public Health Surveill*, vol. 6, p. e19273, May 2020.

[12] X. Zhou, A. Mulay, E. Ferrara, and R. Zafarani, "ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research," *arXiv e-prints*, p. arXiv:2006.05557, June 2020.

[13] "https://developer.twitter.com/en/docs."

[14] "http://www.adsnotizie.it/_testate.asp."

[15] "https://www.bufale.net/."

[16] "https://www.butac.it/."

[17] "https://www.newsguardtech.com/it/."

[18] P. Lucisano and M. E. Piemontese, "Gulpease. una formula per la predizione della difficoltà dei testi in lingua italiana," pp. pp. 57–68, 1988.

[19] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[20] F. Pierri, "The diffusion of mainstream and disinformation news on twitter: The case of italy and france," in *Companion Proceedings of the Web Conference 2020*, WWW '20, (New York, NY, USA), p. 617–622, Association for Computing Machinery, 2020.

[21] K.-C. Yang, C. Torres-Lugo, and F. Menczer, "Prevalence of low-credibility information on twitter during the covid-19 outbreak," *ArXiv*, vol. abs/2004.14484, 2020.

[22] "https://www.tweetbinder.com/blog/twitter-impressions/."

[23] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala, "The covid-19 social media infodemic," 2020.

[24] P. Peñas, R. del Hoyo, J. Vea-Murguía, C. González, and S. Mayo, "Collective knowledge ontology user profiling for twitter – automatic user profiling," in *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, vol. 1, pp. 439–444, 2013.

Penalty	**C**	**Solver**
l1-l2-none	0.01-0.1-1	liblinear-sag-saga

Table 3. Logistic Regression.

Neighbors	**Weights**	**P**
range(1,100,4)	Uniform-Distance	1-2

Table 4. KNN.

Estimators	**Criterion**	**Max Features**
range(10,200,10)	Gini-Entropy	log2-sqrt-None

Table 5. Random Forest.

Criterion	**Max Features**
Gini-Entropy	log2-sqrt-None

Table 6. Decision Tree.

C	**Degree**	**Kernel**	**Gamma**
0.01-0.1-1-2	1-2-3-4-5	linear-poly-rbf-sigmoid	0.01-0.1-1.0

Table 7. SVM.

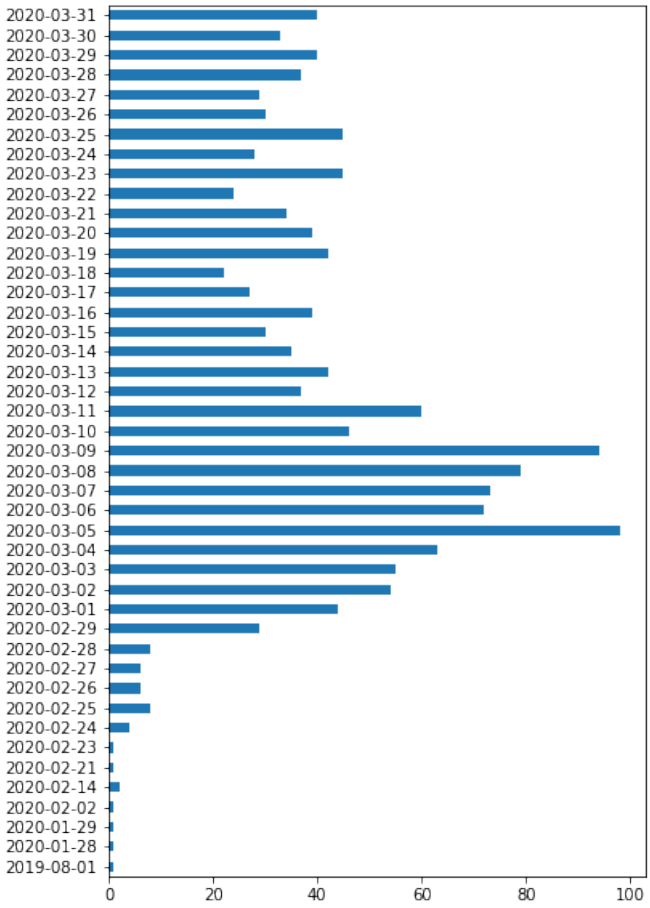
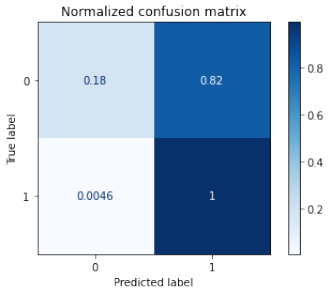
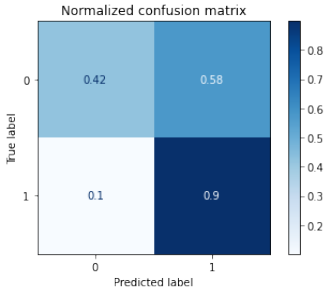
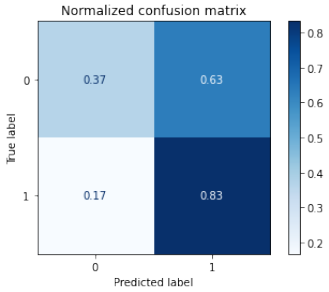
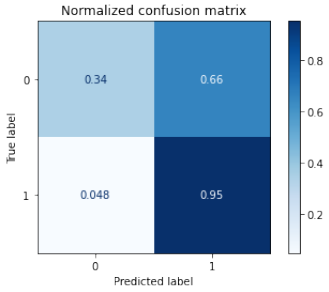


Figure 5. Publish Date.









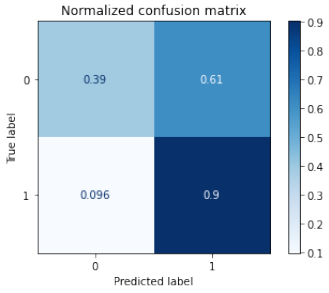
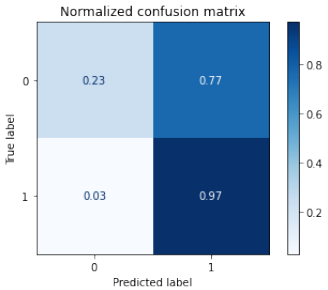
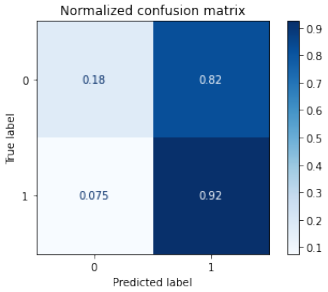
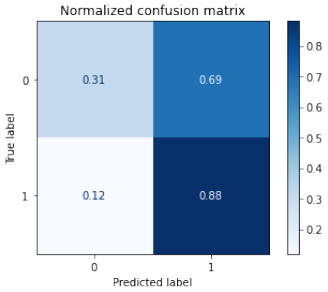
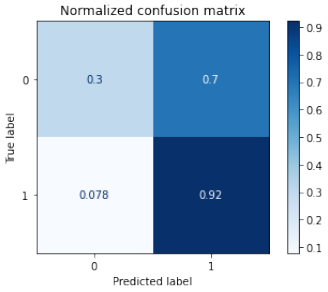


Figure 6. Logistic Regression, KNN, Decision Tree, Random Forest, SVM.









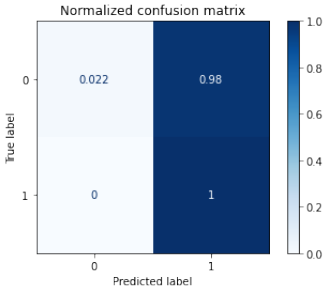


Figure 7. Logistic Regression, KNN, Decision Tree, Random Forest, SVM.

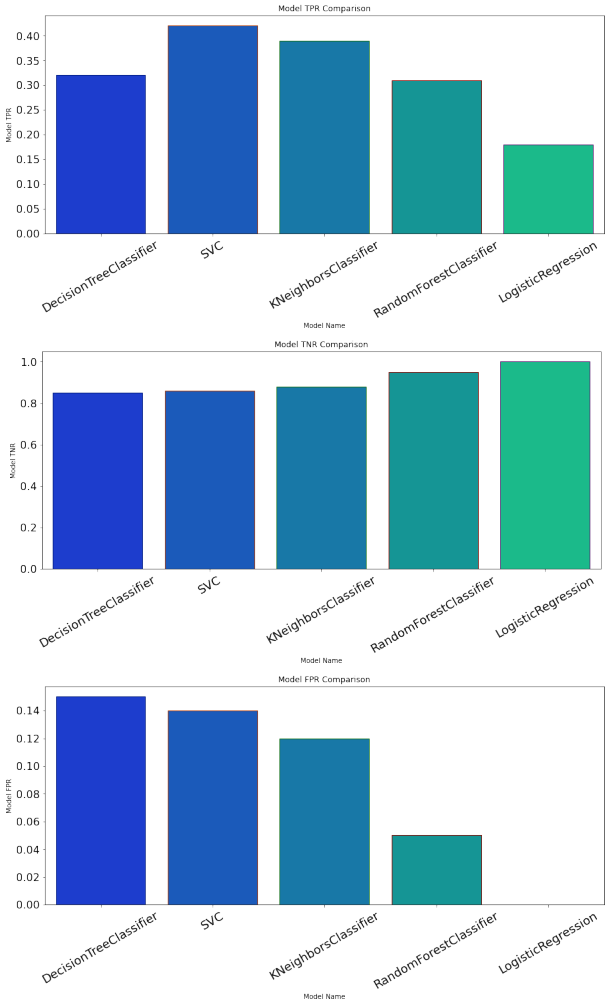


Figure 8. Comparison on stylometry features classifiers.

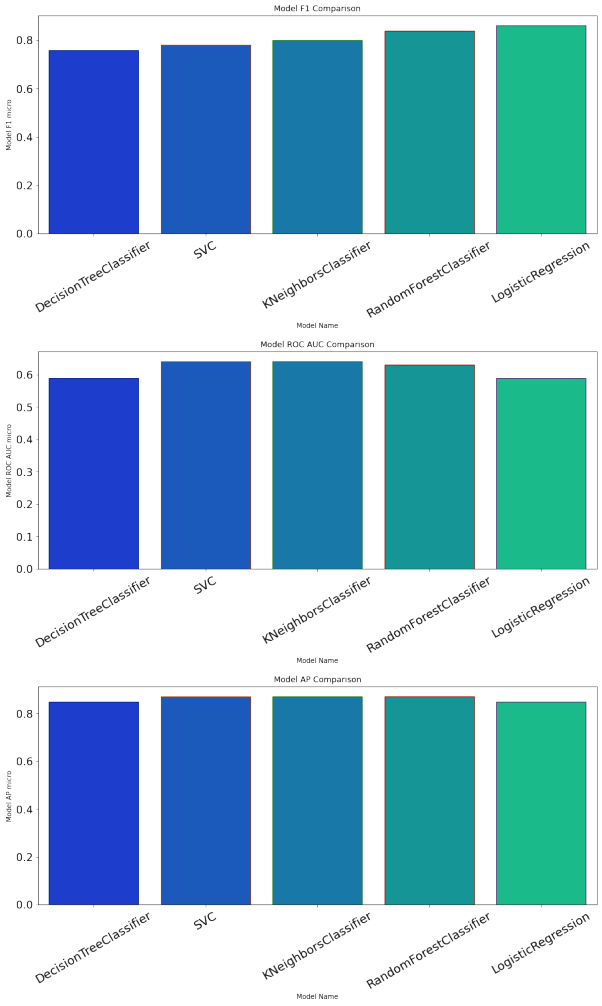


Figure 9. Comparison on stylometry features classifiers.

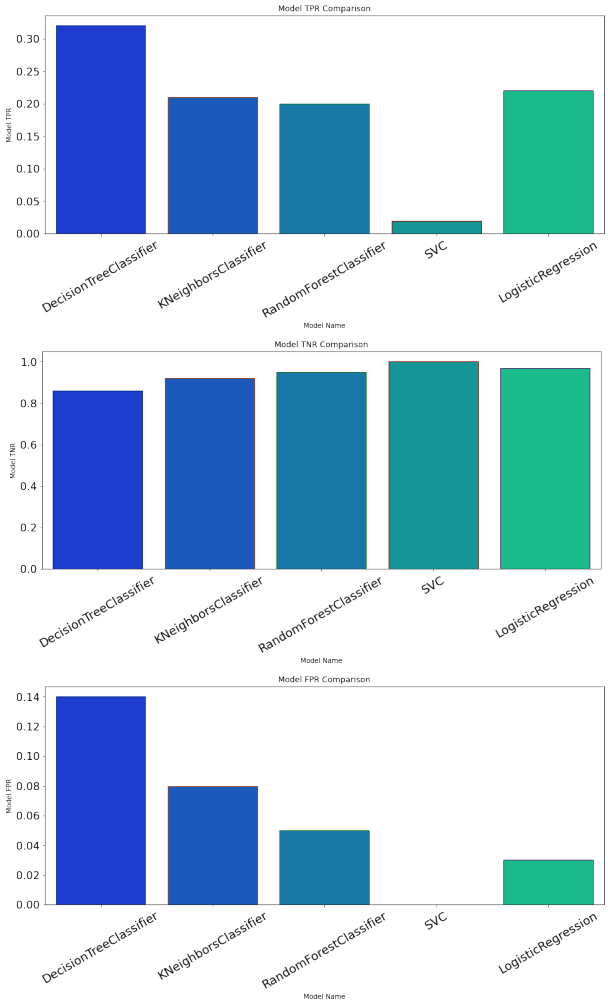


Figure 10. Comparison on TF-IDF features classifiers.

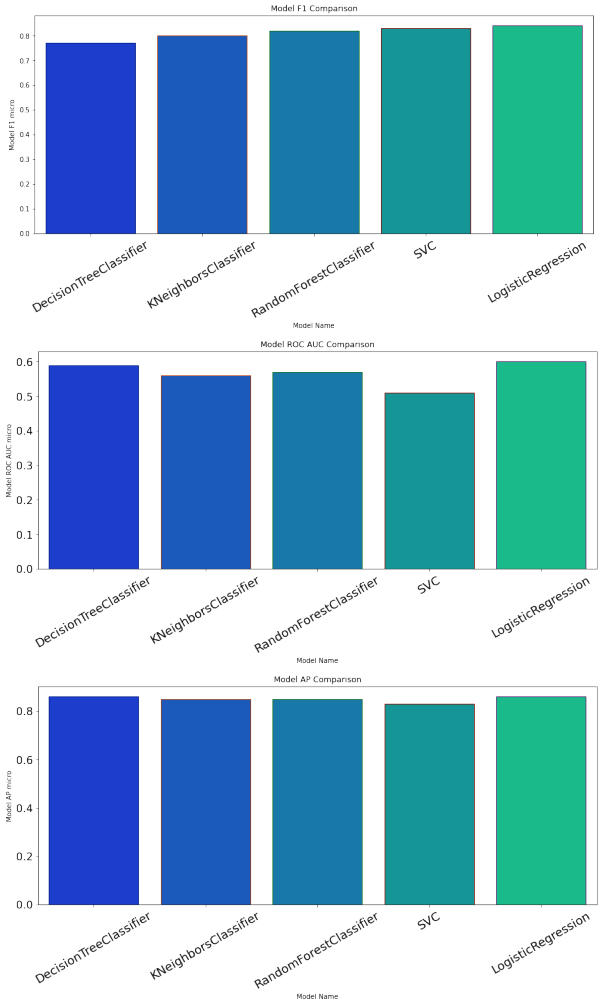


Figure 11. Comparison on TF-IDF features classifiers.

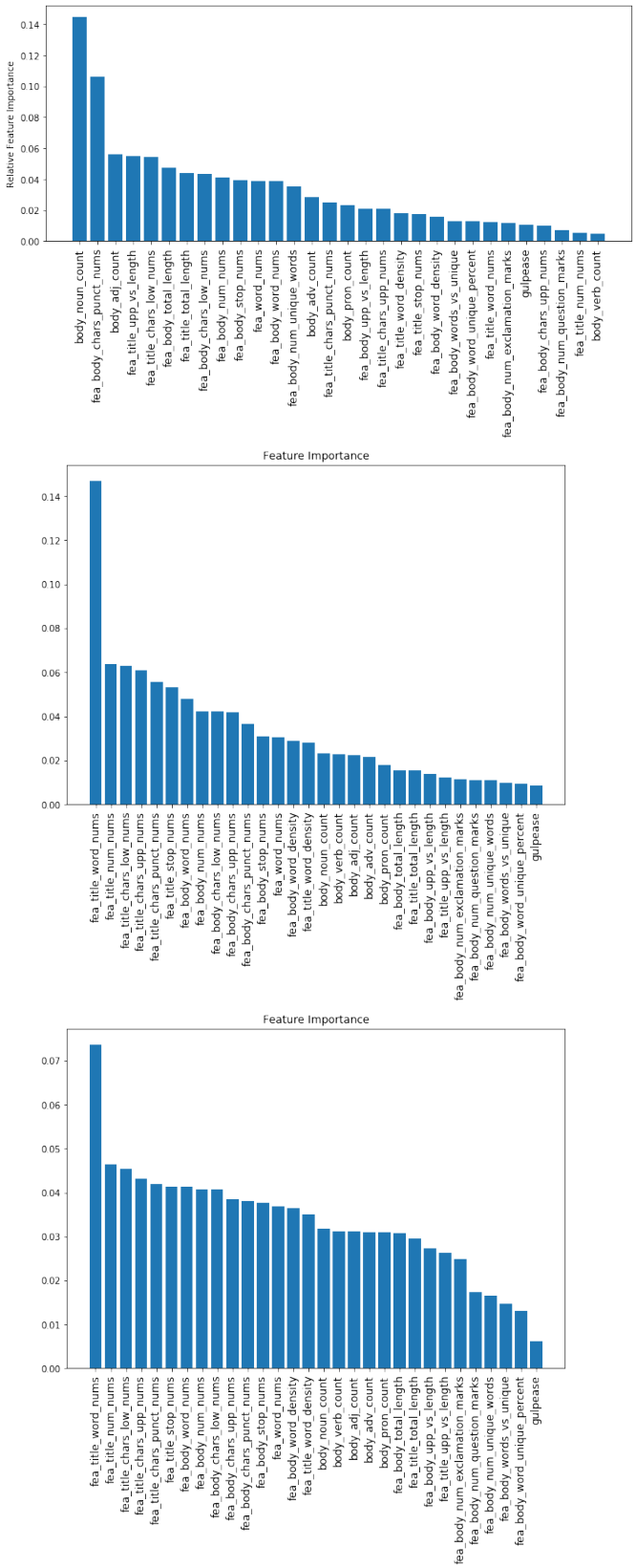


Figure 12. Feature Importances: Logistic Regression, Decision Tree, Random Forest.

Feature Name	Characteristic
fea_title_word_nums	n words in title
fea_title_num_nums	n numerics in title
fea_title_chars_low_nums	n low chars in title
fea_title_chars_upp_nums	n upper chars in title
fea_title_chars_punct_nums	n punctuation in title
fea_title_stop_nums	n stopwords in title
fea_body_word_nums	n words in body
fea_body_num_nums	n numerics in body
fea_body_chars_low_nums	n low chars in body
fea_body_chars_upp_nums	n upper chars in body
fea_body_chars_punct_nums	n puctuation in body
fea_body_stop_nums	n stopwords in body
fea_word_nums	n words in title + body
fea_body_word_density	n chars over n words in body
fea_title_word_density	n chars over n words in title
body_noun_count	n noun in body
body_verb_count	n verb in body
body_adj_count	n adjective in body
body_adv_count	n adverb in body
body_pron_count	n pronoum in body
fea_body_total_length	length of body
fea_title_total_length	length of title
fea_body_upp_vs_length	rate upper char to length in body
fea_title_upp_vs_length	rate upper char to length in title
fea_body_num_exclamation_marks	n ! in body
fea_body_num_question_marks	n ? in body
fea_body_num_unique_words	n of unique words in body
fea_body_words_vs_unique	rate words to unique words in body
fea_body_word_unique_percent	percentage of prev features

Table 8. Features.

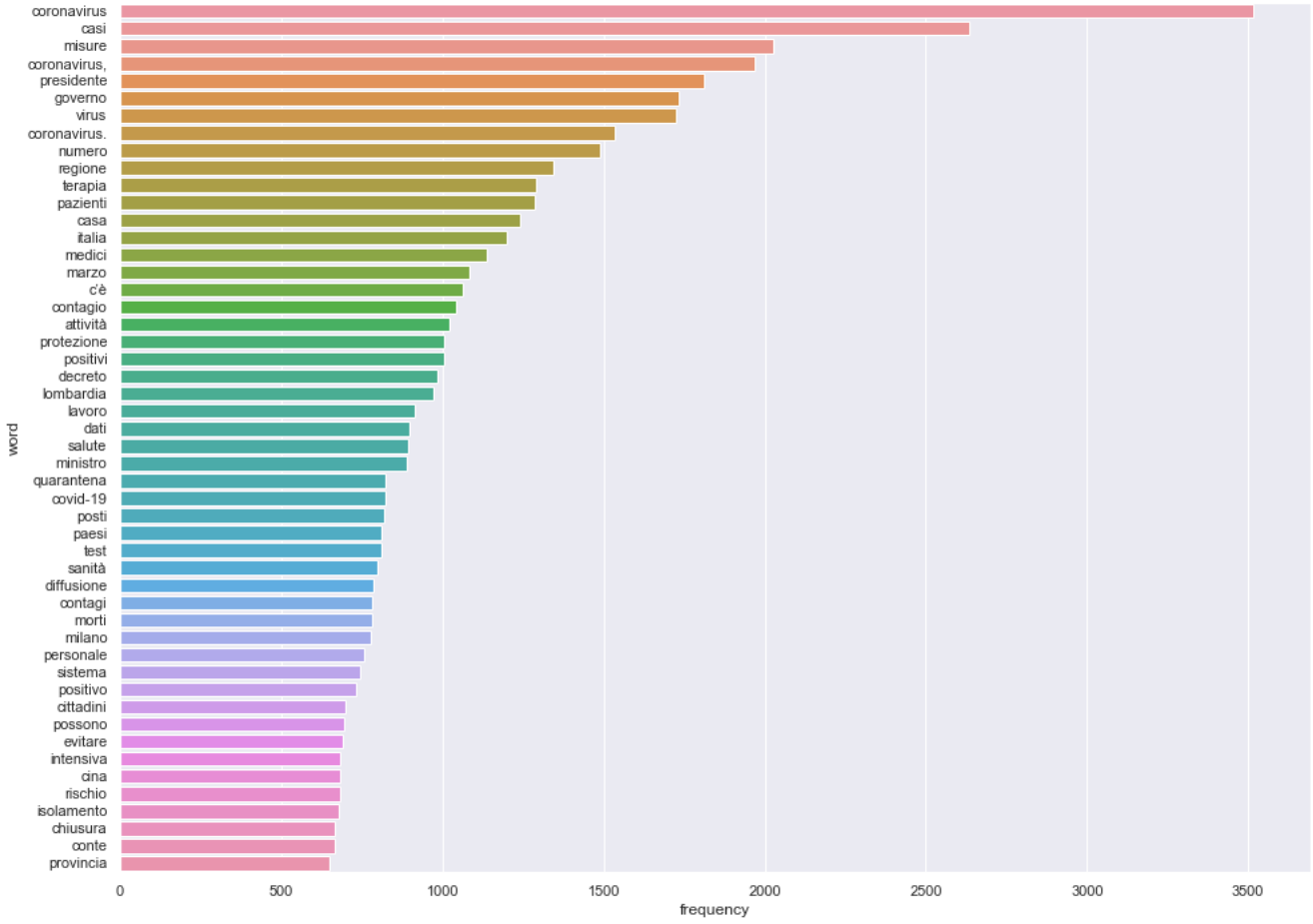


Figure 13. Word Count Bar.