Low Credibility Media on Twitter during Covid-19 Emergency in Italy

Piero Romare

Abstract

Nowadays, misinformation is a known phenomenon, but the diffusion 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 true 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). We perform classification task between the two classes using different Machine Learning models. Lastly, we measure impact and stakeholders from LCM networks based on user engagements and friendship. Our analysis shows that LCM are shared with a ratio of 1:5 to HCM and have a huge potential impact on over 120.000 users at first degree and on over 330.000 users at second degree.

1. Introduction

In online social media, people are no longer just passive readers of news 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 recommended 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 have 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 social media tend to prefer real information over false news, in terms of questions from users to community [4]. The information process is complex, users can generate information either by providing their observations, by bringing relevant knowledge from external sources, or by deriving interpretations [5]. Discussions are in a continuous process and during the start of 2020, the worldwide Social Media debate is more about Covid-19 [6].

Misinformation detection is a major challenge. Covid-19 open some interesting developments come from Google [7] and a collaboration between European Parliament and LaRepubblica [8]. Fake news classification is tested in Facebook posts with Bayes Classifier with 74% of test accuracy [9] and many other studies rely to supervised methods with in scenarios [10, 11, 12, 13]. In [14] authors proposed an Long Short-Term Memory neural network model that is emotionally infused to detect false news. A survey is elaborated in [15] where there is a distinguish between content based, context based and both using different machine and deep learning models. Regarding graph theory and low credibility websites detection, DistrustRank constructs a weighted graph where nodes represent websites, connected by edges based on a minimum similarity between a pair of websites to spot false news domain [16]. Yang et al. build a network where characteristic behavior from accounts which amplify misinformation from the same sources and at fairly similar by measuring the similarity between pairs of those accounts [17].

In this project, we classify news articles, in terms of low credibility (i.e. news come from namely websites which notably produce disinformation) or high credibility (i.e. news come from traditional and mainstream outlets), shared on Twitter in the month of March [18]. 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 [19]. Furthermore, we investigate the impact of LCM on Twitter network building two kinds of graphs and extend [20]: news-based and user-based.

1.1. Contribution

In this project, we make the following contributions:

- We collect Italian scenario tweets on Covid-19;
- We evaluate different machine learning classifiers to discriminate HCM and LCM using two features extraction approaches;

- We give an overall phenomenon and provide the impact of potential impact of LCM on Twitter and the relative stakeholders;

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.

2. Method

In this project, we can divide two macro-areas of analysis: one's concerns binary classification on articles content-based features to discriminate HCM and LCM, the other concerns community network graphs to measure the phenomenon.

2.1. Machine Learning

2.1.1 Data Collection

We randomly collect a dataset of 190.000 Italian tweets [18] from March using Twitter API [21]. We select those tweets which are attached an URL. URLs can be shared as original URL or as shorten URL and, in this last case, we stretch it to obtain the original domain. We select those tweets which have a domain in [22, 23, 24, 25], discarding retweets to to homogenize the degree of depth of the analysis. We scrape all full body text of articles. Every article has label as reliable article in HCM if its URL domain is in [22] (Fig. 1) and as un-reliable articles in LCM if its URL domain are in [23, 24, 25] (Fig. 2). Results show that 2192 articles are reliable and 447 articles are un-reliable (see Appendix Algorithm 1). In order to visualize the dataset, we show a Word Cloud based on word frequencies (Fig. 3) to provide an overview of the contents relating to the most common words used in HCM and LCM articles. In particular, bigger the word represented in the graph the more it is present in the articles (e.g., "emergenza", "coronavirus", "ospedale", "italia", "covid") which reflect the contents. We provide the distribution of publish date of articles scraped (Fig. 4). A first exploration of data shows that the retweet engage are with a mean of 70 per tweet of HCM and 43 per tweet of LCM retweet, and the situation change for what concern favorite engagement, where a mean of 3.5 is for HCM and a mean of 7 for LCM. So, while HCM are more reshare than LCM, LCM are more engage.

2.1.2 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:

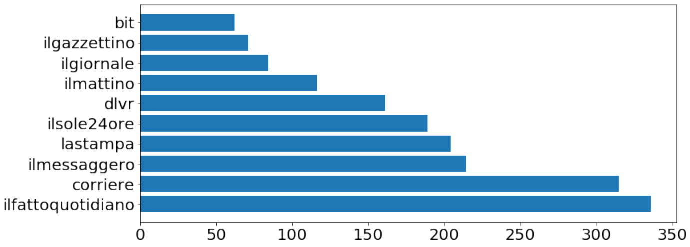

Figure 1. Top 10 Most Common High Media.

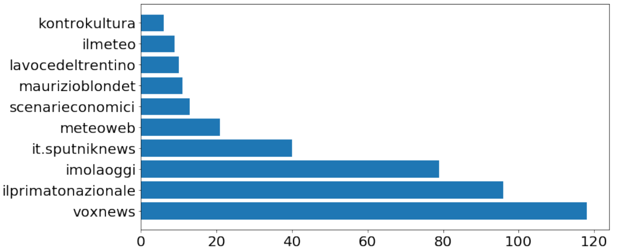

Figure 2. Top 10 Most Common Low Media.

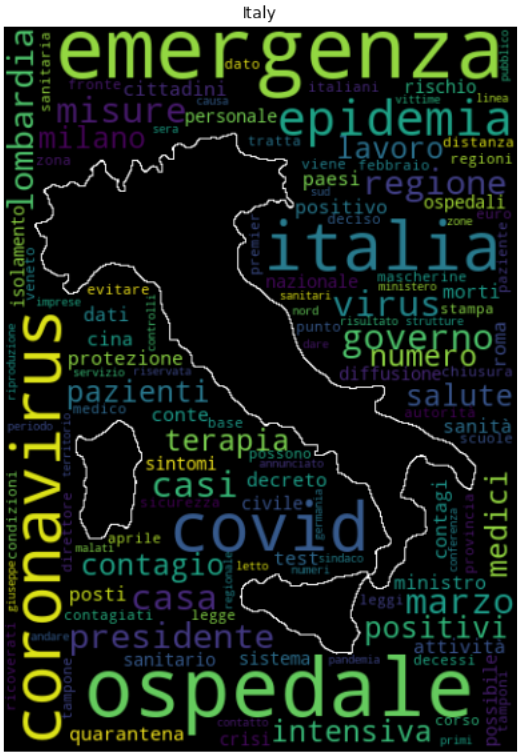

Figure 3. Word Cloud.

1. Stylometry, where we extract content-based features from title and full body text for every articles (Table 7). We add the *Gulpease Index* [26], an Italian readability index calculates in the body text that scores based on how difficult it is to read. We transform features using *StandardScaler*: $z = \frac{(x-u)}{s}$.

2. *Frequency–Inverse Document Frequency (TF-IDF)* on words, 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, in general, *TF-IDF* is technique useful to preprocess features, 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). *TF-IDF* is useful when you might classify different documents topics (e.g., sports vs politics vs science) but, in our case, the documents regarding to a unique topic (e.g., covid-19) in both classes considered. Our goal, using this approach, is to understand if there are linguistic patterns that distinguish HCM from LCM.

2.1.3 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 (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).

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

For F1 evaluation we use a micro average due to the fact that the dataset is unbalanced: the micro weighs each sample equally.

2.2. Graphs

A graph is defined as $G = (V, E)$ where V is the set of vertices or nodes and E is the set of edges. A direct graph has edges, with direction, which connect a node x to another node y or to itself. Directions are represented with an arrow $\rightarrow$ (e.g., $x \rightarrow y$), in other words it suggests a relationship between two nodes.
Here, it follows metrics to evaluate networks:

- *in-degree* with respect to a node means the relations which arrive to the considered node.
- *out-degree* with respect to a node means the relations which start from the considered node;
- A directed graph is called *weakly connected* if replacing all of its directed edges with undirected edges produces a connected graph. In our context, users that are in relationship each other;
- *Density* is basically the proportion of the edges in the graph with respect to all possible edges.

In this context, it is useful define centrality measures to identify stakeholders (e.g., how influential a person, an opinion leader, is within a network) [27]:

- *Degree Centrality* indicates the number of nodes in which the considered node is connect. In our context, for finding very connected individuals, popular individuals, individuals who are likely to hold most information or individuals who can connect with the wider network.
- *Closeness Centrality* indicates the node that can achieve all other nodes in the graph, in other words the easiest node that get all other nodes in the graph. In our context, for finding the individuals who are best placed to influence the entire network most quickly.

Furthermore, we define potential impact as the count of followers of users which share a LCM. Potential impact indicates the number of users to whom an LCM can appear on the bulletin board.

We produce 4 different graphs divided in two classes:

1. News-based: we build a weighted direct graph for both possibility in which the nodes are the users in the filtered dataset of LCM, the edges exist if at least one of the following rules are respected [20, 28]:

   - a is retweeted by b;
   - b is quoted by a;
   - a mentions a;
   - b replies to a.

   We assign the weight based on the count of interactions described in the bullets above iteratively with 1 (i.e. the node a and the node b have two interactions, described above, we assign a weight=2).

2. User-based: we build a direct graphs for both possibility with LCM in which the nodes are the users in the filtered dataset of LCM, the edges exist if the following rule is respected:

   - a is following a'.

Model	**TPR**	**TNR**	**F1**
Logistic	0.18	1.00	0.86
KNN	0.42	0.90	0.82
Decision	0.37	0.83	0.75
Random	**0.34**	**0.95**	**0.85**
SVM	0.39	0.90	0.82

Table 1. Models Evaluations based on stylometry features.

Model	**TPR**	**TNR**	**F1**
Logistic	**0.48**	**0.98**	**0.89**
KNN	0.86	0.28	0.38
Decision	0.67	0.91	0.87
Random	**0.60**	**0.98**	**0.91**
SVM	0.50	0.97	0.89

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

3. Results

3.1. Machine Learning

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 6) with scoring on *precision* and 5-Fold cross validation. We consider features importances for Logistic Regression which resulting be that the number of nouns in the body articles and the number of punctuation in the body articles are the most important, Decision Tree where the number of words in the title is the most important and Random Forest where the importance have a similar rank of Decision Tree features (Fig. 5). The results with stylometry features are provided in Table 1. As regards instead *TF-IDF* features, we provide words importances for Logistic Regression for class HCM and LCM in (Fig. 6) where can be seen that words as ”cinese” and ”clandestini” are useful to the classifier to assign label HCM to articles and ”università”, ”lavoro” and ”studio” to LCM articles. The results with *TF-IDF* features are provided in Table 2. In our context, 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. In our case, we choose, as mentioned before, a micro average for the metrics and we obtain the same value for F1. Overall, we obtain that *Logistic Regression* and *Random Forest* as best estimator for our task with stylometry features. *Logistic Regression* classify correctly 18% LCM and 100% for what concern HCM, while *Random Forest* achieve 0.34 of TPR and 0.95 of TNR. They both obtain the highest F1 micro score. For what concern the models trained with *TF-IDF* features, the results achieve higher scores to others with stylometry features, we evaluate also here the *Random Forest* as best estimator due the fact that it show a TPR of 0.60, misclassify LCM as HCM in 2% of cases and perform a F1 score of 0.91. *Logistic Regression* is the second best classifier which correctly classify LCM in 48% of cases, HCM in 98% and achieve a F1 score of 0.89. From our analysis, we can conclude that *TF-IDF* features perform better than stylometry, considering that the evaluations are performed on same models.

Metric	
N Nodes	1052
N Edges	1215
Avg Degree	1.15
N Weakly Component	35
Nodes without out-degree	941
Nodes without in-degree	89
Density	0.001
Max in-degree	17
Max out-degree	294

Table 3. News Network Evaluations.

3.2. Graphs

In this paragraph, we propose the experimental results of the Twitter interaction network related to quantify the community and its relationships of users who share LCM. We express stakeholders with the respective user id following the guidelines in Terms of Service (ToS) of Twitter.

3.2.1 News-based

In Table 3, we show the metrics of the overall News Network. Here, 1052 participate in LCM network, 250 are single users who share a LCM, while 802 are users which engage with at least one of the LCM present in the dataset. There are 1215 connection between these users. 35 users interacts with LCM each other. PAY ATTENTION HERE **(941 of users are not involved in the engage described in section 2.2 (news-based), while 89 users have no visibility with their posts or have not share a post)**.

In Table 4, we show the metrics of the News Network with weight > 1 to highlight stronger connection between users and engaged users: 105 participate in LCM network with at least two rules respected in section 2.2 (News-based). There are 91 connection between these more active users with respect to previous network unweighted. 17 of these users follow each other. 84 of users are not involved in the engage, while 17 users have no visibility with their posts or have not share a post. These results are based on a filter from the News-based network to recognize which are the users the mostly participate in LCM engage.

Metric	
N Nodes	105
N Edges	91
Avg Degree	0.87
N Weakly Component	17
Nodes without out-degree	84
Nodes without in-degree	17
Density	0.008
Max in-degree	2
Max out-degree	34

Table 4. News Network Evaluations (weight $>$ 1).

Metric	
N Nodes	250
N Edges	3446
Avg Degree	13.78
N Weakly Component	48
Nodes without out-degree	92
Nodes without in-degree	75
Density	0.06
Max in-degree	59
Max out-degree	98

Table 5. User Network Evaluations.

We provide the user id which achieve the highest metric score considered in the relative network:

1. News Network stakeholders: the user 1683455144 has a degree centrality of 0.28; the user 2695501 has a closeness centrality of 0.02; Users who share and interact with LCM have a potential impact on over 335,000 users and each of them are reached with an average of 3 times. We can consider these values as the number of users reached in a second degree of the news network, but in our case, without considering the users, and their related followers, engaged in a like interaction on the tweets containing LCM.

2. News Network stakeholders, with weight $>$ 1: the user with id 1683455144 has a degree centrality of 0.33; the user with id 407747764 has a closeness centrality of 0.02; In the news network with weight $>$ 1 the potential impact is on over 83.000 users and each of them are reached with an average of 2 times. The assumption about like interaction is valid also for news network with weight $>$ 1.

3.2.2 User-based

In Table 5, we show the metrics of the Users Network. Here, 250 unique users participate in LCM network. There are 3446 connection between these users, it means that on average every user following other 14 users already present in the network. 48 users follow each other, in other words, 20% of users has a strong participation and informally we can say that they form a community. 92 users do not following none of users present in the network, while 75 users are not followed by other users in the network. We provide the user id which achieve the highest metric score considered in the relative network:

1. User Network stakeholders: the user 1121911847840571393 has a degree centrality of 0.59; the user 4394960301 has a closeness centrality of 0.38; The users which share LCM have a potential impact on 120.000 users and each of them are reached on average from 4 different users.

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 0.91 and correctly classify a HCM in the 98% of the cases and correctly in 60% of times the LCM. Also *Logistic Regression* give us a F1 score of 0.89 and in the 48% of times misclassify LCM as HCM. Indeed, the graphs show us the stakeholders of the diffusion of LCM. With news-based, we provide an overview of interactions between users that share LCM and users which engage with them. Furthermore, we are able to estimate the potential impact of diffusion, at the first degree, of LCM in user-based: 250 unique users which shared a LCM potentially address it to over 120.000 users. For what concern the potential impact of LCM diffusion at the second degree, we consider the potential impact of news network that achieve over 335.000 users.

4.1. Future Work

As future work, we first plain to extend our Machine Learning exploration with context-based features. Thus, possible directions: *SIS* epidemiological model to estimate *Echo Chambers* [29], Politics Profiling of user-based on which news or common link they share [30], it could be also use for weighting directed graphs to estimate *Cognitive Opinion Dynamics* [31] 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] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences," in *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, (USA), p. 129–136, Association for Computational Linguistics, 2003.

[6] 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.

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

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

[9] 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.

[10] S. Dori-Hacohen and J. Allan, "Detecting controversy on the web," in *Proceedings of the 22nd ACM International Conference on Information Knowledge Management*, CIKM '13, (New York, NY, USA), p. 1845–1848, Association for Computing Machinery, 2013.

[11] B. D. Horne and S. Adali, "This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news," 2017.

[12] S. Kumar, R. West, and J. Leskovec, "Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes," in *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, (Republic and Canton of Geneva, CHE), p. 591–602, International World Wide Web Conferences Steering Committee, 2016.

[13] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum, "Credibility assessment of textual claims on the web," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, (New York, NY, USA), p. 2173–2178, Association for Computing Machinery, 2016.

[14] 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.

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

[16] V. Woloszyn and W. Nejdl, "Distrustrank: Spotting false news domains," in *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, (New York, NY, USA), p. 221–228, Association for Computing Machinery, 2018.

[17] 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.

[18] 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.

[19] 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.

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

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

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

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

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

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

[27] "https://cambridge-intelligence.com/keylines-faqs-social-network-analysis/."

[28] M. Cinelli, S. Cresci, A. Galeazzi, W. Quattrociocchi, and M. Tesconi, "The limited reach of fake news on twitter during 2019 european elections," *PLOS ONE*, vol. 15, pp. 1–13, 06 2020.

[29] 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.

[30] 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.

[31] D. Vilone, F. Giardini, M. Paolucci, and R. Conte, "Reducing individuals' risk sensitiveness can promote positive and non-alarmist views about catastrophic events in an agent-based simulation," 2016.

5. Appendix

Params	Logistic Regression	KNN	Random Forest	Decision Tree	SVM
Penalty	l1-l2-none				
C	0.01-0.1-1				
Solver	liblinear-sag-saga				
Neighbors		range(1,100,4)			
Weights		Uniform-Distance			
P		1-2			
Estimators			range(10,200,10)		
Criterion			Gini-Entropy	Gini-Entropy	
Max Features			log2-sqrt-none	log2-sqrt-none	
Degree					1-2-3-4-5
Kernel					linear-poly-rbf-sigmoid
Gamma					0.01-0.1-1

Table 6. Grid Search Features.

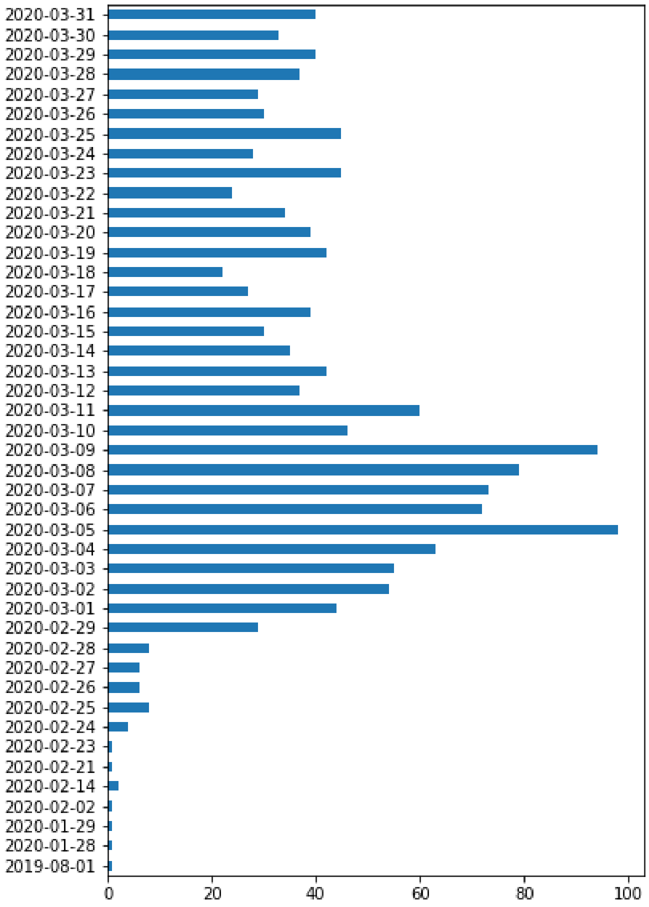

Figure 4. Publish Date.

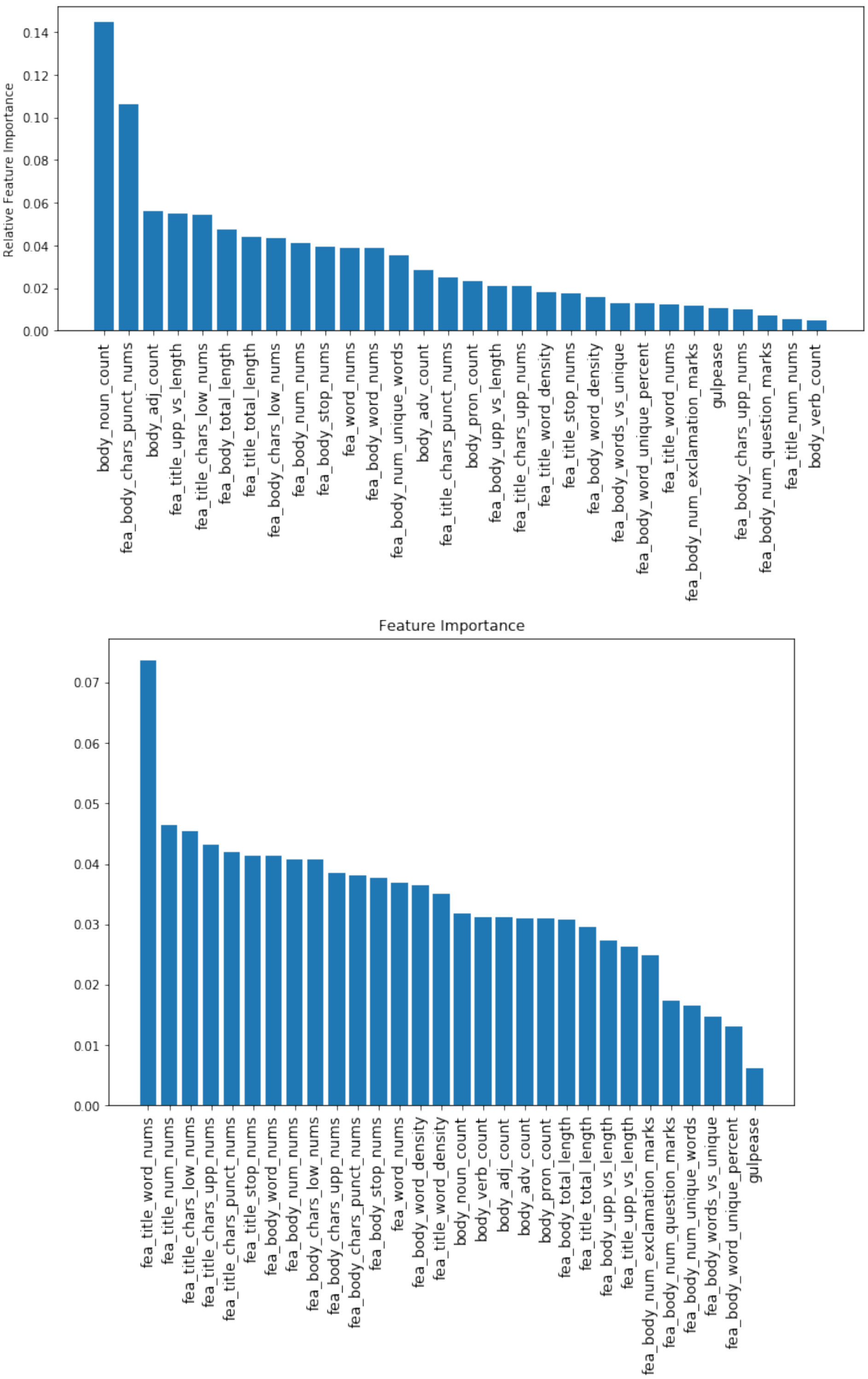

Figure 5. Logistic Regression and Random Forest - Stylometry Features Importance.

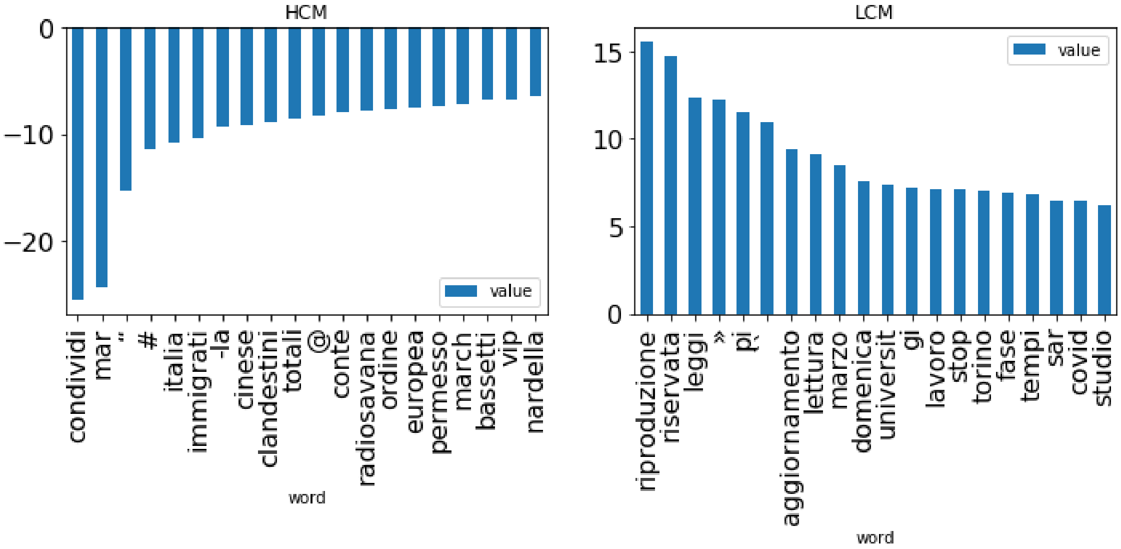

Figure 6. Logistic Regression TF-IDF Words Importance.

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 7. Features.

Algorithm 1

```
HCM = set of HCM domain
LCM = set of LCM domain
Shorten domains = set of URL shortener services
for tweet in tweets set do
    tweet link = tweet.entities.expanded_url
    tweet link = Regular Expression to extract domain
    if tweet link then
        if tweet link in Shorten domains then
            tweet link = unshorten(tweet link)
        end if
        if tweet link in HCM then
            tweetHCM.extend(tweet link)
        end if
        if tweet link in LCM then
            tweetLCM.extend(tweet link)
        end if
    else
        pass
    end if
end for
return tweetHCM, tweetLCM
```