*Machine Learning and Artificial Intelligence*
*J.-L. Kim (Ed.)*

*doi:10.3233/FAIA230772*

# Satiric Content Detection Through Linguistic Features

Angelo GAETA [a], Francesco ORCIUOLI [a,1] and Antonella PASCUZZO [a]
[a] *Department of Management and Innovation Systems*
*Università degli Studi di Salerno*
*Via Giovanni Paolo II, 132 - 84084 Fisciano (SA), Italy*

**Abstract.** In the age of Information Disorder, Satire is one of its phenomena mainly occurring in the context of social media. Satire represents an interesting study subject given that it can be easily confused with further forms of the disorder. The present work proposes and evaluates a set of linguistic features to build classifiers able to distinguish satires from other textual contents. The adopted features are firstly identified within the scientific literature and, secondly, ranked and filtered by means of the Information Gain index. Several experimentation activities show good performance for the aforementioned classifiers and an acceptable ability to generalize for the models trained with such features.

**Keywords.** Satire detection, Information disorder, Machine learning, Linguistic features

## 1. Introduction

The coming of the Internet and Social Media has resulted in changes in the way in which people are linked with each other and the information is created, produced, and distributed. Moreover, the speed at which a large amount of information is transmitted has completely changed. Nowadays, more and more researchers approached the phenomena of Information Disorder in the digitally-connected world. The Council of Europe has presented a Conceptual Framework [1] to better explain such phenomena by looking at typologies, elements, and phases. Concerning typologies, two different dimensions can be considered: the truthness of the content and the intention to harm. Along the aforementioned dimensions, it is possible to classify the content as mis-information, dis-information, or mal-information. From the computational viewpoint, the approaches to studying Information Disorder can be classified into two main groups: i) *network structure* approach able to analyze spreading, motivation, and intent, ii) *machine learning algorithm and text mining-based* approaches which allow detecting several forms of information disorder through the extraction of specific features. Some of these forms are, for instance, fake news, hoaxes, hate speech, conspiracy, propaganda, etc. In this context, the present work mainly focuses on *satires* (satiric textual content) which are defined

[1]Corresponding Author: Francesco ORCIUOLI, Department of Management and Innovation Systems, Università degli Studi di Salerno, Via Giovanni Paolo II, 132 - 84084 Fisciano (SA), Italy. E-mail: forciuoli@unisa.it.

by [2] as "a phenomenon where humor and irony are employed to criticize and ridicule someone or something". Satire is typically classified as a mis-information type of information disorder given that it is, in general, fake and has no intention to harm. According to the need for better characterizing satiric content, the authors of [3] try to distinguish satires from hoaxes (fake content with the intent to harm) and propaganda (fake or real content with or without intent to harm). Further works try to describe, in more detail, the characteristic of satiric content. In particular, the authors of [4] affirm that satire is characterized by: controversial or sensitive issues, aggressive language - negative emotions and tone - for entertainment purposes, a shorter form with respect to true news even if the words that can be found are more complex, a language not so clear since they are not written by professional journalists, based on imagination and figurative language such as metaphors, similes, personifications, idioms, etc. Moreover, in [5], three language dimensions are considered to better understand satires: the use of first-person singular, which is a proclamation of one's ownership of statements, the more negative words reflect negative emotions, and the use of exclusive words which emphasize a cognitive complexity. In such a context, the present work assesses the effectiveness of linguistic-based features, extracted from textual content, of supporting the automatic detection of satires by means of the application of well known machine learning methods. Through a series of experiments, it is demonstrated that such kind of features (especially those related to the readability of textual content) effectively support the building of good classification models having an acceptable ability of generalization. The above aspects represent the main contribution to the field of automated detection of information disorder.

## 2. Related works

From a computational perspective, satire was not much studied in the past. Recently, researchers have begun to look at it from a different computational perspective. For instance, authors of [6] adopt an emotion-based approach and those of [7] proposed a method based on stylometry. Only a few authors adopt linguistic-based approaches to accomplish the satire detection task. For example, the authors of [8] consider seven classes of features to analyze: frequency, ambiguity, part-of-speech (POS), synonyms, sentiments, characters, and slang words. In [9], the authors just consider the headlines of the considered articles by extracting profanity and slang words and using Name Entity Recognition (NER). Moreover, Yang, Mukherjee, and Gragut [10] identified four main categories of features able to guarantee an acceptable characterization of satiric content: Writing-Stylistic, Readability, Structural, and Psycho-linguistic. Several works, like [6,8,11,12], leverage on Doc2Vect or Term Frequency Inverse Document Frequency (TF-IDF) to construct features able to train classifiers for satires. Such approaches, on one side, provide good performance but, on the other side, fail to characterize satire in a human-understandable and explainable way. In the present work, the idea is to extract (from textual content) and integrate features introduced by the analyses conducted in both [8] and [10] and, subsequently, evaluate the performance of several machine learning approaches when supporting by such features. The experimentation activities demonstrated good results (better than those reported by the existing literature) and an acceptable and promising generalization ability of the trained model. The obtained results are also aligned with the theoretical analyses introduced in Section 1.

## 3. Feature engineering

In order to construct the dataset for supporting the building of classification models for executing the satire detection tasks, it was needed to construct an initial set of suitable features. The identified features were mainly extracted from textual content and aligned to the literature results reported in Section 1. The analysis of the literature produced a set of feature categories. Subsequently, tools able to extract the identified features from the text were selected. Therefore, the final set of features was determined by considering the specific functionalities of the selected tools and the features considered suitable by the specialized literature. The result of this process is reported in Table 1.

**Table 1.** Set of text-based features.

| Category | Features | Cit. | Tools |
|---|---|---|---|
| **Writing-Stylistic** | Distribution of Pronoun, Determiners, Preposition, Adverbs, Adjectives, Verbs, Conjunctions, Interjections, and Negation. | [8] [13] | Python Script by the Authors and LIWC Software results. |
| **Psycho-linguistic** | Summary Dimensions, Linguistic, Basic and Expanded Dictionary. | [13] | LIWC Software. |
| **Structural** | Maximum and Average length of words, distribution of Upper, Lower characters, and punctuations (quotes, exclamation and join line ("-"). | [12] | Python Script by the authors and LIWC Software results. |
| **Readability** | Flesch Reading Ease, Gunning Fog, Dale Chall, Automated Readability, Coleman Liau, Linsear Write, SPACHE, and Entropy value. | [12] | Python Script by the authors using the Py-Readability-Metrics Package. |
| **Sentiment/Emotions** | Scores of Positive, Neutral, Negative, Compound, Polarity and Subjectivity. | [14] | Python Script by the authors using VADER, EmoLex, and TextBlob. |
| **Slang and Profanity** | Distribution of the Slang and Profanity. | [8] | Python Script by the authors using a Slang dict and Prof. list. |
| **Ambiguity** | Maximum and Average number of Synsets, Frequency Gap as the difference between these two measures. | [8] | Python Script by the authors using WordNet. |

In particular, the Linguistic Inquiry and Word Count[2](LIWC) software was employed (with LIWC-22 English Dictionary) to extract *Psycho-linguistic* and support the extraction of *Structural* and *Writing-Stylistic* features. Moreover, VADER Sentiment Analysis Tool[3], TextBlob library[4] and EmoLex[5] were used to extract sentiment and emotions from textual content. Lastly, *Py-Readability-Metrics Package*[6] was adopted to pro-

[2]https://www.liwc.app
[3]https://github.com/cjhutto/vaderSentiment
[4]https://textblob.readthedocs.io
[5]https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
[6]https://pypi.org/project/py-readability-metrics/

duce scores for the readability level of a text by using the most popular metrics and the *WordNet*[7] library was employed to measure the ambiguity level within a text.

## 4. Experimentation and evaluation

Several experiments have been conducted to select the most important features from the initial set discussed in Section 3, evaluate the performances of different machine learning algorithms, and assess the generalization capabilities of the selected features. Except for the feature engineering phase (realized through the implementation of Python scripts), the rest of the employed data pipeline was realized by using Orange3[8].

### *4.1. Dataset construction*

The present work uses three datasets for the experimentation phase. The dataset A is obtained by processing the News Headlines Dataset for Sarcasm Detection [15] composed of non-satirical and satirical headlines with the relative URL and a Boolean labeling feature `is_sarcastic`. Since the authors wanted to investigate the linguistic patterns of long text, it has been scraped the full-text content of such articles started from the available URLs. Once performed in this first phase, long textual content was processed as indicated in Section 3 in order to obtain 99 total features. Subsequently, all the 96 numerical features were normalized in the range [0, 1] by means of a min-max approach. The same processing task was executed for datasets B and C also exploited to assess the ability to generalize of the trained model. Dataset B (including a total of 13.804 rows) was obtained by a set of news articles from HuffPost[9] (labeled as non-satiric) and satirical news articles from The Onion[10] (labeled as satiric). Lastly, dataset C (including 1.152 rows of satire and hate speech articles - where the latter are treated as non-satirical) was obtained by relabeling the dataset Fake News Corpus[11] that is very problematic given that it was originally labeled by-hands and provides several rows that could be associated to different target classes (of course only one class is selected).

### *4.2. Experiments and results*

Three main experiments were conducted with the aim of measuring the performance of classification models built through five well known machine learning methods[12][16] and by exploiting the linguistic-based features considered in Section 3. The first one focused on the application of the aforementioned methods on the whole set of features separately on datasets A, B and C. The second one consisted in applying the same methods on dataset A by using top-15 most relevant features resulting from the Information Gain index (see Table 2).

[7] https://www.nltk.org/howto/wordnet.html
[8] https://orangedatamining.com/
[9] https://www.huffpost.com/
[10] https://www.theonion.com/
[11] https://github.com/several27/FakeNewsCorpus
[12] Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM), AdaBoost (AB), Logistic Regression (LR)

**Table 2.** Results of the first two experiments.

| | 15-top features on dataset A | | | | | All features on dataset A | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **RF** | **LR** | **AB** | **NB** | **SVM** | **RF** | **LR** | **AB** | **NB** | **SVM** |
| **AUC** | 0.998 | 0.985 | 0.970 | 0.960 | 0.918 | 0.999 | 0.996 | 0.975 | 0.962 | 0.987 |
| **CA** | 0.980 | 0.929 | 0.972 | 0.885 | 0.853 | 0.985 | 0.973 | 0.975 | 0.887 | 0.934 |
| **F1** | 0.980 | 0.930 | 0.972 | 0.886 | 0.853 | 0.985 | 0.973 | 0.975 | 0.889 | 0.934 |
| **Precision** | 0.981 | 0.932 | 0.972 | 0.898 | 0.853 | 0.986 | 0.973 | 0.975 | 0.906 | 0.936 |
| **Recall** | 0.980 | 0.929 | 0.972 | 0.885 | 0.853 | 0.985 | 0.973 | 0.975 | 0.887 | 0.934 |
| | **All features on dataset B** | | | | | **All features on dataset C** | | | | |
| | **RF** | **LR** | **AB** | **NB** | **SVM** | **RF** | **LR** | **AB** | **NB** | **SVM** |
| **AUC** | 0.997 | 0.989 | 0.939 | 0.956 | 0.991 | 0.959 | 0.891 | 0.764 | 0.907 | 0.892 |
| **CA** | 0.971 | 0.968 | 0.938 | 0.908 | 0.949 | 0.891 | 0.809 | 0.787 | 0.765 | 0.800 |
| **F1** | 0.971 | 0.968 | 0.938 | 0.908 | 0.949 | 0.892 | 0.812 | 0.787 | 0.771 | 0.799 |
| **Precision** | 0.971 | 0.969 | 0.938 | 0.912 | 0.949 | 0.894 | 0.826 | 0.786 | 0.822 | 0.799 |
| **Recall** | 0.971 | 0.968 | 0.938 | 0.908 | 0.949 | 0.891 | 0.809 | 0.787 | 0.765 | 0.800 |

The third one is performed (on dataset A) by using the above methods on subsets of features corresponding to the categories introduced in Table 1. All the three experiments were conducted by using the 80% of the dataset for the training task and the remaining 20% for the test. Table 2 reports the results of the first two experiments. In particular all the obtained accuracy scores (CA) are higher (except for some methods applied on dataset C) than the best literature results for satire detection provided by [11] that achieves an accuracy of 0.834 on the same data of dataset A. Lastly, the third experiment demonstrates that Readability features provides the best support for satire classification when using the Logistic Regression (accuracy of 0.920), Random Forest (accuracy of 0.975), or AdaBoost (accuracy of 0.958).

### *4.3. Ability to generalize*

The performance of the classifiers trained (on dataset A) and tested (on datasets B and C) significantly improves when exploiting 25-top features in place of 15-top features (both obtained through the Information Gain index). Naive Bayes demonstrates an acceptable ability to generalize on different test datasets as the number of features grows. In fact, for dataset B, Naive Bayes achieves AUC of 0.600 (top-15) and 0.674 (top-25) while the best results are obtained by Logistic Regression with AUC score of 0.746 (top-25). For dataset C, the best results are provided by Random Forest with AUC score of 0.602. Lastly, when applying the Recursive Feature Elimination [17], a set of 29 features, coherent with the theoretic literature, was obtained. In this case, Logistic Regression provided AUC of 0.887 (test over dataset B) and Naive Bayes obtained AUC of 0.652 (test over dataset C).

## 5. Discussion and final remarks

This paper proposes and evaluates the adoption of linguistic-based features, extracted from text, to support the detection of satires through the application of machine learning. The work demonstrates good results when the classifiers are trained and tested over two different parts (80%-20%) of the same dataset. An acceptable attitude to generalize is shown when the trained models are tested on different datasets where scores are sensibly lower but grow as the number of features increases. The selected 15-top and 25-top features are also explainable through the theoretic literature results in Table 1.

## Acknowledgement

This work was partially supported by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU.

## References

[1] Wardle C, Dias D. Information disorder: Toward and interdisciplinary framework for research and policy. Council of Europe; 2017.
[2] Colletta L. Political Satire and Postmodern Irony in the Age of Stephen Colbert and Jon Stewart. The Journal of Popular Culture. 2009 Oct;42:856-74.
[3] Rashkin H, Choi E, Jang JY, Volkova S, Choi Y. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics; 2017. p. 2931-7.
[4] García-Díaz JA, Valencia-García R. Compilation and evaluation of the Spanish SatiCorpus 2021 for satire identification using linguistic features and transformers. Complex & Intelligent Systems. 2022;8:1723-36.
[5] Newman ML, Pennebaker JW, Berry DS, Richards JM. Lying Words: Predicting Deception from Linguistic Styles. Personality and Social Psychology Bulletin. 2003;29(5):665-75.
[6] Thu PP, New N. Implementation of emotional features on satire detection. In: 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD); 2017. p. 149-54.
[7] Reganti A, Maheshwari T, Das A, Cambria E. Open Secrets and Wrong Rights: Automatic Satire Detection in English Text. In: Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. New York, NY, USA: ACM; 2017. p. 291-4.
[8] Barbieri F, Ronzano F, Saggion H. Is This Tweet Satirical? A Computational Approach for Satire Detection in Spanish. Procesamiento del Lenguaje Natural. 2015.
[9] Burfoot C, Baldwin T. Automatic Satire Detection: Are You Having a Laugh? In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. Suntec, Singapore: Association for Computational Linguistics; 2009. p. 161-4.
[10] Yang F, Mukherjee A, Dragut E. Satirical News Detection and Analysis using Attention Mechanism and Linguistic Features. CoRR. 2017;1709.01189.
[11] Ahmad T, Akhtar H, Chopra A, Akhtar MW. Satire Detection from Web Documents Using Machine Learning Methods. In: 2014 International Conference on Soft Computing and Machine Intelligence; 2014. p. 102-5.
[12] Jain MK, Gopalani D, Meena YK, Kumar R. Machine Learning based Fake News Detection using linguistic features and word vector features. In: 2020 IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON); 2020. p. 1-6.
[13] Boyd RL, Ashokkumar A, Seraj S, Pennebaker JW. The development and psychometric properties of LIWC-22. Austin, TX: University of Texas at Austin. 2022:1-47.
[14] Frain A, Wubben S. SatiricLR: a language resource of satirical news articles. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16); 2016. p. 4137-40.
[15] Misra R, Arora P. Sarcasm Detection using News Headlines Dataset. AI Open. 2023:13-8.
[16] Ray S. A Quick Review of Machine Learning Algorithms. In: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon); 2019. p. 35-9.
[17] Chen Xw, Jeong JC. Enhanced recursive feature elimination. In: Sixth International Conference on Machine Learning and Applications (ICMLA 2007); 2007. p. 429-35.