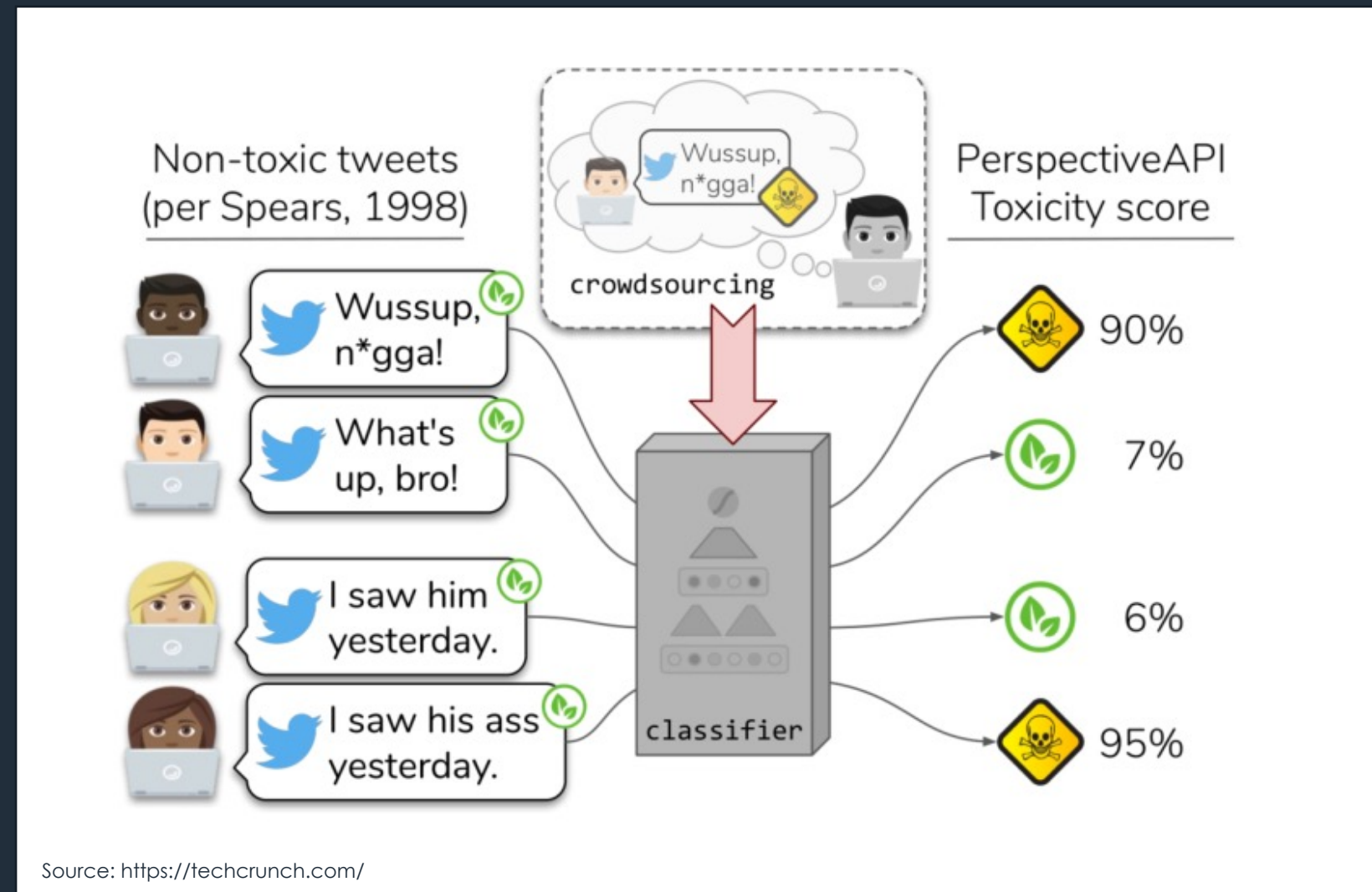


# Automated Harmful Content Detection Using Grammar- Focused Representations of Text Data

Daria Stetsenko (NASK PIB),  
Inez Okulska (NASK PIB),  
Kinga Głębicka (NASK PIB)



# Are these sentences offensive?

You should know women's sports are a joke.

Offensive

Neutral

All mental illnesses are awful and must be treated.

Offensive

Neutral

Men and women are not equal. Irrational contrary belief and policy only result in mounting failure.

Offensive

Neutral

You look like someone who would do an electric wheelchair race with Stephen Hawking.

Offensive

Neutral

# What do we consider as offensive language?



Help Center > General > The Twitter Rules

## The Twitter Rules

---

Hateful conduct includes language that dehumanizes others on the basis of religion or caste. In March 2020 the hateful conduct policy was expanded to also include race, ethnicity, or national origin. Following this context, hate speech can be defined as an abusive speech that targets specific group characteristics, such as gender, religion, or ethnicity.

[https://blog.twitter.com/en\\_us/topics/company/2019/hatefulconductupdate.html](https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.html)

# Facebook Community Standards

The Facebook Community Standards outline what is and isn't allowed on Facebook.

Hate speech is a direct attack against people on the basis of protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease.

Community standards. [https://www.facebook.com/communitystandards/hate\\_speech](https://www.facebook.com/communitystandards/hate_speech)



Hate speech detection using static BERT embeddings. Gaurav Rajput, Narinder Singh punn, Sanjay Kumar Sonbhadra, and Sonali Agarwal.

# A tweet is offensive if it:

- > uses a sexist or racial slur;
- > attacks a minority;
- > seeks to silence a minority;
- > criticizes a minority (without a well-founded argument);
- > promotes, but does not directly use, hate speech or violent crime;
- > criticizes a minority and uses a straw man argument;
- > blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims;
- > shows support of problematic hash tags;
- > negatively stereotypes a minority;
- > defends xenophobia or sexism;
- > contains a screen name that is offensive.



Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. Zeerak Waseem and Dirk Hovy.

# Why should we care?

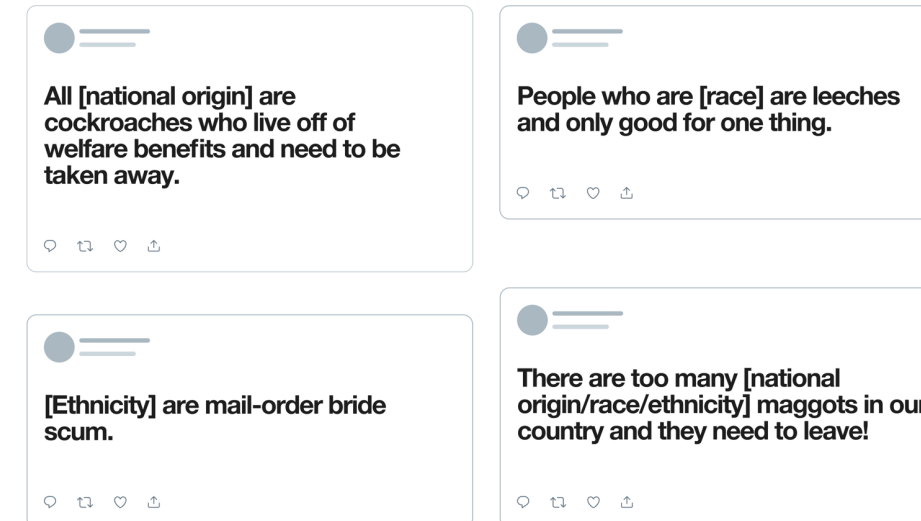


HAMAS PALESTINE  
@b4ng\_yus

Lets kill jews and kill them for fun

#killjews

7/20/14, 8:05 AM







# Is hate speech easy detectable?

Model	F1-Score	Accuracy	Precision	Recall	Specificity
CNN + Attention + FT + GV	74.41	75.15	74.92	74.35	80.35
<b>CNN + Attention + static BE</b>	77.52	77.96	77.89	77.69	79.62
CNN + LSTM + GV	72.13	72.94	73.47	72.4	76.65
<b>CNN + LSTM + static BE</b>	76.04	76.66	77.20	76.18	79.43
LSTM + FT + GV	72.85	73.43	73.37	72.97	76.44
<b>LSTM + static BE</b>	79.08	79.36	79.38	79.37	79.49
BiLSTM + FT + GV	76.85	77.45	77.99	77.10	79.66
<b>BiLSTM + static BE</b>	<b>79.71</b>	<b>80.15</b>	<b>80.37</b>	<b>79.76</b>	<b>83.03</b>
BiLSTM + Attention + FT	76.80	77.34	77.76	77.00	79.63
<b>BiLSTM + Attention+static BE</b>	78.52	79.16	79.67	78.58	83.00
<b>GRU + static BE</b>	77.91	78.36	78.59	78.18	79.47
BERT	78.83	76.64	79.17	78.43	74.31

Bold model names represent static BERT embedding variants of the models

Bold values represent the highest value of any metric among all models



“Hate speech detection using static BERT embeddings.” Gaurav Rajput, Narinder Singh punn, Sanjay Kumar Sonbhadra, and Sonali Agarwal



# What is StyloMetrix?

spaCy

The metrics are:



## INTERPRETABLE

each metric represents an aspect of linguistic knowledge



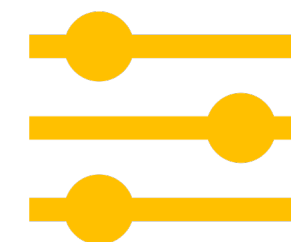
## NORMALIZED

metrics express number of occurrences of given feature per number of tokens in text, which lets us escape scaling effect in texts of different lengths



## REPRODUCIBLE

values of metrics can be recalculated or even counted manually giving always the same output. The representation doesn't depend on any random factor or seeding



## CUSTOMIZABLE

if your needs exceed the scope of built-in metrics, create your own! Don't forget to share your work and contribute to the community of StyloMetrix!



Inez Okulska [inez.okulska@nask.pl](mailto:inez.okulska@nask.pl) | Anna Zawadzka [anna.zawadzka@nask.pl](mailto:anna.zawadzka@nask.pl)

# The most distinct syntactic features to detect hate speech

- G\_ADJ\_COMPARATIVE

SY\_QUESTION

SY\_SPECIAL\_QUESTION

SY\_EXCLAMATION

SY\_IMPERATIVE

SY\_SUBORD\_SENT

SY\_SUBORD\_SENT\_PUNCT

SY\_COORD\_SENT

SY\_SENT\_START\_ADV







SY\_SENT\_START\_ADJ

POS\_PREP

L\_PROPER\_NAME

L\_PERSONAL\_NAME

L\_PUNCT\_COL

	Low Intensity	High Intensity
Costliness	<div><div></div><div><div>Hate Speech Warner</div><div>@hate_suspension</div></div></div> <p>The user @account you follow was suspended, and I suspect that this was because of hateful language. If you continue to use hate speech, you might get suspended temporarily. @your_account</p>	<div><div></div><div><div>Hate Speech Warner of High Suspension Costs</div><div>@suspension_cost</div></div></div> <p>The user @account you follow was suspended, and I suspect that this was because of hateful language. If you continue to use hate speech, you might lose your posts, friends and followers, and not get your account back. @your_account</p>
Legitimacy	<div><div></div><div><div>Hate Speech Warner</div><div>@WarnerHate</div></div></div> <p>The user @account you follow was suspended, and I suspect that this was because of hateful language. Your tweets bother me – you should stop using hate speech to avoid suspension. @your_account</p>	<div><div></div><div><div>Hate Speech Warner due to Suspension Risk</div><div>@ban_warner</div></div></div> <p>The user @account you follow was suspended, and I suspect that this was because of hateful language. I understand that you have every right to express yourself but please keep in mind that using hate speech can get you suspended. @your_account</p>
Credibility	<div><div></div><div><div>Just a human</div><div>@basic_person_12</div></div></div> <p>The user @account you follow was suspended, I suspect that this was because of hateful language. Twitter suspends thousands of users each month. I don't know much about Twitter, but my guess is that you might get suspended too if you continue to use hate speech. @your_account</p>	<div><div></div><div><div>Hate Speech Detector Based on Data</div><div>@expert_on_hate</div></div></div> <p>User @account you follow was suspended, I suspect that this was due to hateful language. Twitter suspends thousands of users each month. I am a professional researcher who studies suspensions due to hate speech. My model says that you might also get suspended. @your_account</p>

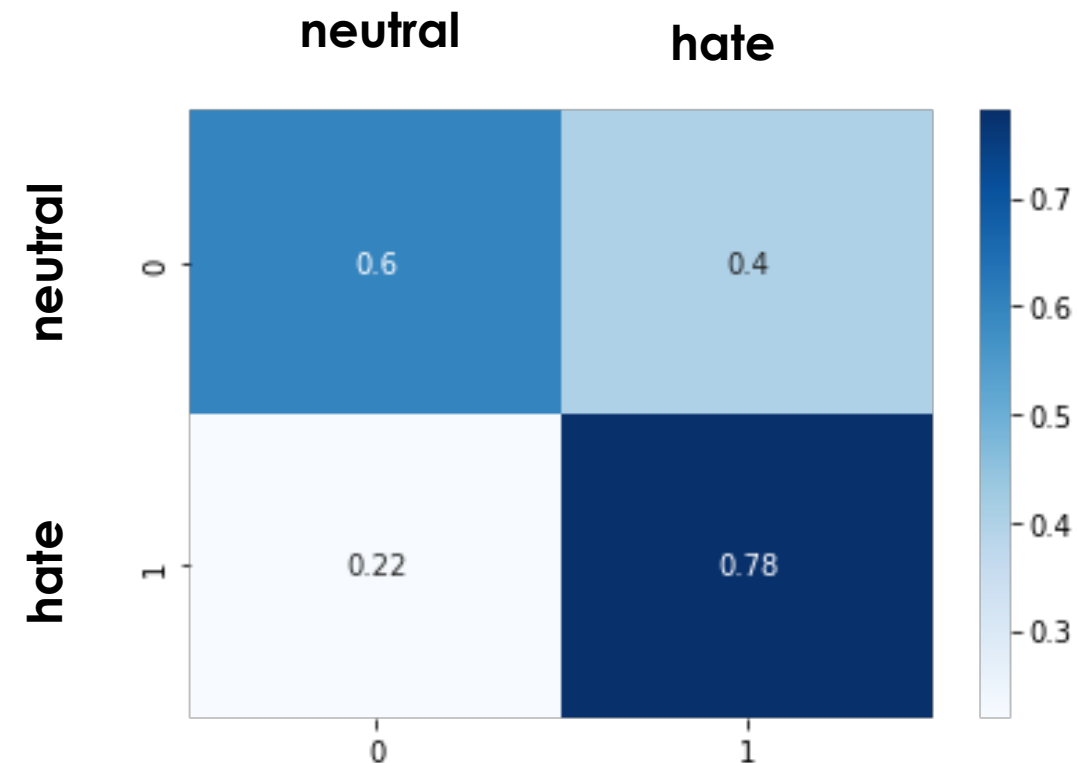
# StyloMetrix Vectors on Dynamically Generated Hate Speech Dataset (English language)

40463 tweets (54% hate, 46% neutral)

Voting Classifier:

- Linear Regression
- Random Forest Classifier
- AdaBoost

**78%** for hate



# Models on Dynamically Generated Hate Speech Dataset (English language)

**LSTM (GLOVE)**

**LSTM (GLOVE) probabilities +  
StyloMetrix -> VotingClassifier**

**hate**

**78%**

**81%**

**neutral**

**74%**

**76%**

# Models on Dynamically Generated Hate Speech Dataset (English language)

pre-trained RoBERTa

pre-trained RoBERTa probabilities  
+ StyloMetrix -> VotingClassifier

hate

**80%**

**81%**

neutral

**77%**

**77%**

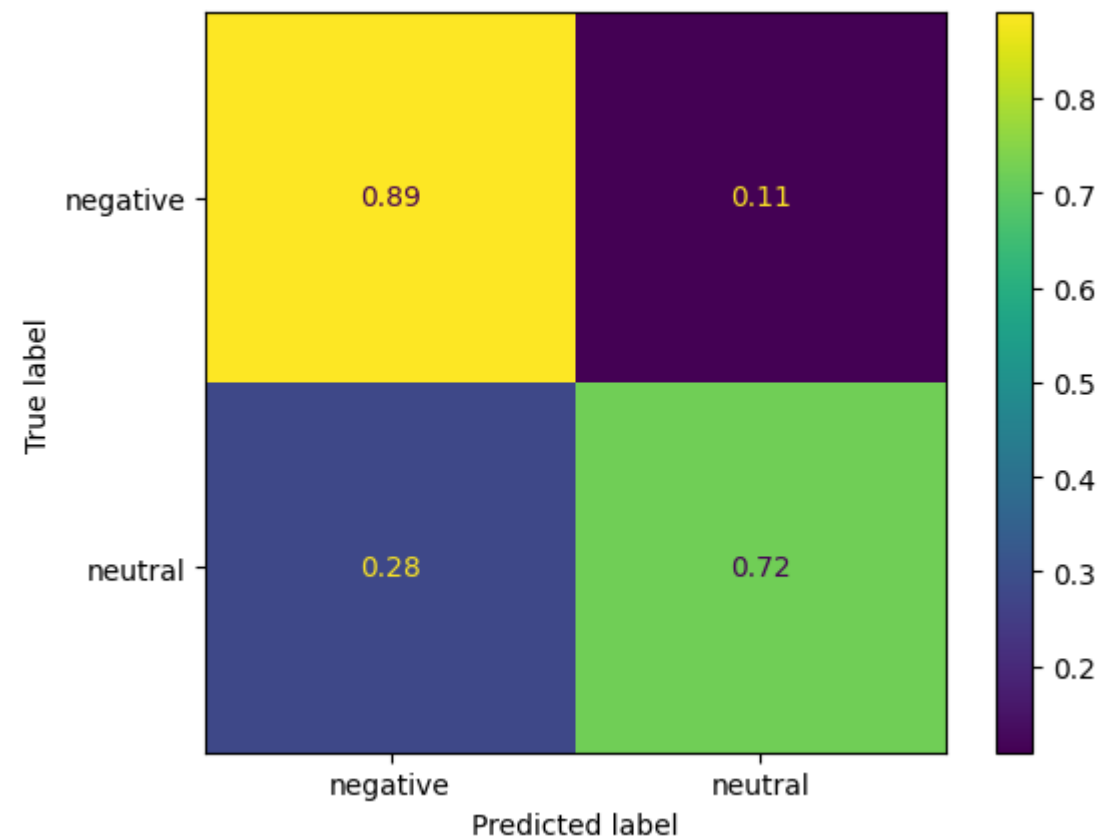


# StyloMetrix Vectors on Wykop.pl comments (Polish language)

## Voting Classifier:

- Linear Regression
- Random Forest Classifier
- AdaBoost

**89%** for hate

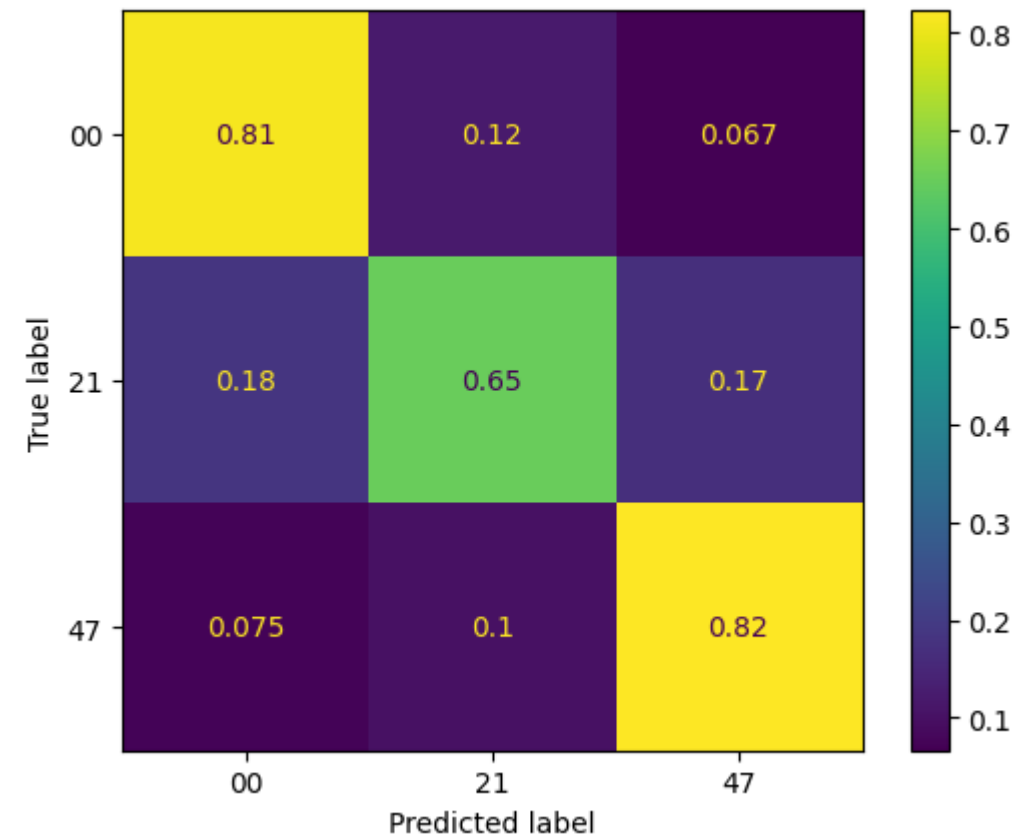


# StyloMetrix Vectors on Wykop.pl comments (Polish language)

Voting Classifier:

- Linear Regression
- Random Forest Classifier
- AdaBoost

**73,5%**  
for hate classes



# StyloMetrix Vectors on Wykop.pl comments (Polish language)



	fine-tuned RoBERTa	fine-tuned RoBERTa probabilities + StyloMetrix -> VotingClassifier
hate classes	87,5%	89,5%
neutral	99%	99%

# StyloMetrix – open source and waiting for you!

github.com/ZILiAT-NASK/StyloMetrix

☰ README.md

## StyloMetrix



Zakład Inżynierii Lingwistycznej i Analityki Tekstu, NASK PIB

### 📌 Quick

- 💡 Stylometry tool in beta version for **Polish** and **English** language, distributed as a Python package
- 💡 [Tutorial notebook](#)
- 💡 List of built-in metrics for [Polish](#), [English](#)
- 💡 [Helper functions and extensions](#)

### 📄 Citation

Please cite [this article](#) when referring to StyloMetrix:

```
Okulska, I., & Zawadzka, A. Styles with Benefits. The StyloMetrix Vectors for Stylistic and Semantic Text Cla
```

### 🔔 About

StyloMetrix is a tool for creating **text representations** as **StyloMetrix vectors**. Each metric in vector quantifies a

**[github.com/ZILiAT-NASK/StyloMetrix](https://github.com/ZILiAT-NASK/StyloMetrix)**

**Thank you!**