Annotation of Multiword Expressions for Hate Speech Identification in the Domain of Sports over Social Media

Panagiotis Grigoriadis (Aristotle University of Thessaloniki)

1st UniDive training school

Chișinău, Moldova

8-12 July 2024

1. Introduction

"Multiword Expressions (MWEs) are lexical items characterized by lexical, syntactic, semantic, pragmatic or statistical idiosyncrasies" (Giouli, Foufi & Fotopoulou, 2019: 125). For this reason, they pose challenges to humans and machines alike. MWEs are a pervasive cross-linguistic phenomenon. They are very often present in texts that propagate hate speech, thus leading to hate crimes which exceed the narrow virtual boundaries (Williams et al., 2020; Perifanos & Goutsos, 2021). They express intense sentimental states of hate and intolerance. Manifestation forms of hate speech include threatening, abuse, harassment and other types of motivation that can lead to embracement of violent practices (Parent, Gobble & Rochlen, 2019). Following examples by Fotopoulou & Giouli (2015): 72 constitute evidence of MWEs encrypting possible hate speech:

1. Ο Νίκος έβγαλε τη Μαρία από τη μέση.
   O Nikos evjale ti Maria apo ti mesi
   The$_{SG.}$ Nikos$_{SG.NOM.}$ took$_{3SG.}$ the Maria$_{SG.ACC.}$ from the middle$_{SG.ACC.}$
   lit. The Nikos took the Maria from the middle.
   "Nikos$_{SG.NOM.}$ took$_{3SG}$ Maria$_{SG.ACC}$. out of the way$_{SG.ACC.}$."
2. Η Μαρία βγήκε από τη μέση.
   I Maria vgike apo ti mesi.
   The$_{SG.}$ Maria$_{SG.NOM.}$ was taken$_{3SG}$ from the middle$_{SG.ACC.}$
   lit. The Maria was taken from the middle.
   "Maria$_{SG.NOM.}$ was taken$_{3SG.}$ out of the way$_{SG.ACC.}$."

What makes the situation even more problematic is the fact that hate speech producers are capable of finding ways to trespass strict community regulations regarding appropriation of postings' content in social media, as imposed by Artificial Intelligence (AI) and Natural Language Processing (NLP) (Waseem & Hovy, 2016; Waseem, 2016). This explains why MWEs resonate with Computational Linguistics. Corpora and tools have been developed since the early 80s in an attempt to achieve a systematic representation of MWEs (Church & Hanks, 1990; Dagan & Church, 1994). The need to take an approach to hate speech detection that exceeds word boundaries arises from the fact that hate speech is not always confined to a single word with an explicitly negative / hateful meaning but may be extended to expressions or texts of larger length (Poletto et al., 2020)

2. Aim of the Project

As mentioned above, MWEs are of great interest for Computational Linguistics and at the same time pose a great challenge. The aim of the current project is to investigate Multiword Expressions (MWEs) which appear in Twitter texts referring to hate speech that produces sport violence. Furthermore, it is wished to study MWEs expressing hate speech in two languages, namely Greek and German. Our study on hate speech is focused on the domain of sports, and particularly on social media texts with an emphasis on Twitter.

Hate speech exhibits increasing tendencies in public speech and especially in social networks like Facebook and Instagram, leading to extension of hate crimes' frequency (Williams et al., 2020; Perifanos & Goutsos, 2021). As far as sport violence is concerned, hate speech consists quite often of MWEs expressing intense sentimental states of rage and intolerance. Earlier research on hate speech has been extended to sociolinguistic level which argues for existence of non-linguistic symbols that provide evidence for hate expressions in various social domains (Jaki & De Smedt, 2019). Forming a collective corpus with all types of hate speech available presupposes consideration of non-linguistic symbols along with MWEs and other linguistic material (Poletto et al., 2017; Pereira-Kohatsu et al., 2019).

A contrastive comparison between Greek and German will be conducted by taking Twitter corpora into account, a new genre besides newswire texts that is now covered in PARSEME (Savary et al., 2018; Savary et al., 2023). Afterwards, a collective hate speech corpus will be proposed for the languages examined with hate speech MWEs such as *take out of the way*, *punch in the face*, *stab to death* etc.

3. Proposal: Resource Description

After sufficient information on hate speech has been collected in the examined languages (Greek and German), a dataset for the automatic identification of hate speech will be developed from various sources. Twitter constitutes the primary resource: the corpus sets derived from Twitter will be comparable with respect to size, time span, topic, and textual genre covered. Moreover, sentences from other parallel corpora such as Opus Corpus (Tiedemann & Nygaard, 2004) and Parallel Global Voices (Prokopidis, Papavassiliou & Piperidis, 2016) may also be included. Annotation will be at two-levels. All data collected will be pre-processed at the level of syntax (dependency annotation). Next, all MWEs (whether indicating hate speech or not) will be annotated in each of the examined languages. It is then intended to use the annotation schema, where MWEs will be grouped according to the categories proposed by PARSEME (Savary et al., 2018; Savary et al., 2023).

This is a planned project that focuses on MWEs expressing hate speech in the field of sports, with Greek and German as examined languages, following the example of related research conducted in the past. Baider & Constantinou (2020) have for instance researched irony speech that appears in social media in Greek and Greek Cypriot. On the other hand, Lekea & Karampelas (2018) have proposed a methodology that helps detect hate speech in Greek which may encrypt terroristic evidence. Open questions related to the project which could be addressed during the brainstorming hackathon might be: (i) how MWE identification helps hate speech detection, and (ii) cross-lingual comparisons of MWEs.

4. Expected Outcomes

By participating in UniDive Training School, it is hoped that the necessary skills in the domain of Computational Linguistics will be acquired in order to make annotation at the level of Dependencies and of MWEs as systematic and informative as possible. This will hopefully contribute to detecting hate speech forms that surpass simple word lexical units. Last but not least, fruitful feedback and discussion on the project's methodology is desirable to shed light on potential weaknesses and strengthen the final outcomes of the project. Thus, the project will be presented in a poster should I be selected to participate.

Keywords

Multilingual Corpus, Twitter Corpus, Annotation of Multiword Expressions, Hate Speech Identification

Selected Literature

Baider, F. & Constantinou, M. (2020). Covert hate speech: A contrastive study of Greek and Greek Cypriot online discussions with an emphasis on irony. *Journal of Language Aggression and Conflict*, *8*(2), 262-287.

Church, K. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22-29.

Dagan, I. & Church, K. (1994). Termight: Identifying and translating technical terminology. In *Fourth Conference on Applied Natural Language Processing* (pp.34-40).

Fotopoulou, A. & Giouli, V. (2015). MWEs: support/light verb constructions vs. fixed expressions in Modern Greek and French. In *Workshop on Multiword units in machine translation and translation technology* (pp. 69-73).

Giouli, V., Foufi, V. & Fotopoulou, A. (2019). Annotating Greek VMWEs in running text: a piece of cake or looking for a needle in a haystack? In *CONFERENCE ON GREEK LINGUISTICS* (pp.125-134).

Jaki, S. & De Smidt, T. (2019). Right-wing German hate speech on Twitter: Analysis and automatic detection. *arXiv preprint arXiv:1910.07518.*

Lekea, I.K. & Karampelas, P. (2018). Detecting hate speech within the terrorist argument: a greek caase. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 1084-1091).

Parent, M.C., Gobble, T.D. & Rochlen, A. (2019). Social media behavior, toxic masculinity, and depression. *Psychology of Men & Masculinities*, *20*(3), 277.

Perifanos, K. & Goutsos, D. (2021). Multimodal hate speech detection in Greek social media. *Multimodal Technologies and Interaction*, *5*(7), 34.

Waseem, Z. (2016). Are you a racist or am I seeing things? Annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science* (pp. 138-142).

Waseem, Z. & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop* (pp. 88-93).

Williams, M.L., Bumap, P., Javed, A., Liu, H. & Ozalp, S. (2020). Hate in the machine: Anti-Black and Anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, *60*(1), 93-117.