Automatic classification using deep learning of hate speech posted on the Internet

Nicolas Zampieri

Supervised by Irina Illina and Dominique Fohr

May 20, 2021

PhD Introduction: one definition of Hate Speech

" Communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion or other characteristic. "
[Nockeby, 2000][1]

[1] Encyclopedia of the American Constitution.

PhD Introduction: hate speech detection task

Classify a tweet as **hateful** or **non-hateful**

- Difficulties:
  - No uniform definition of hate speech
  - Hate may be implied
  - Low resources of annotated hate corpus
- Difficulties of hate speech detection in tweets:
  - Tweet may be grammatically incorrect
  - The abbreviations and slangs may be numerous
- Example of tweet:
  75 million more turks abe to travel within the EU ,lets get out of EU NHS Schools overwhelmed already sorry no more ,immigrants please

- Motivation:
  - We think that some **features** can be useful for Automatic Hate Speech Detection (HSD) system based on **deep learning**
- Goals:
  - Study the **impact of different features** on an HSD system
  - Build a system based on **multi-features** approach for HSD

Multiword Expression Features for Automatic Hate Speech Detection[2]

Nicolas Zampieri, Irina Illina, Dominique Fohr

University of Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France

[2]Paper accepted at NLDB2021 conference: http://nldb2021.sb.dfki.de/

Summary

Introduction: multiword expression

Multiword expression (MWE) is a group of words that present idiomatic and compositional meanings
MWE examples with categories:

- *witness protection program* (Nominal)
- *give a crap* (Verbal idiom)
- *get out* (Full verb-particle construction)
- *have fun* (Full light verb construction)
- *thank you* (Discourse)

Introduction: MWE identification task

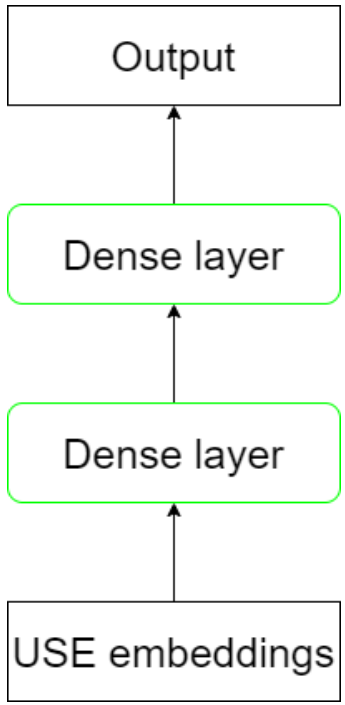
- Automatic MWE identification and text tagging in terms of MWEs are difficult tasks
- There are no state-of-the-art approaches for MWE identification in tweets
- A specific MWE identification system is required to parse MWEs Twitter corpora
- MWE identification system for a tweet corpus is **a complex task**
- We adopt a **lexicon-based approach** to annotate our corpora in terms of MWEs

Methodology

- Our goal is to study **the impact of MWE features** on HSD system
- We combine **sentence-level** feature and **word-level** feature
- We build a model based on neural networks

Methodology: sentence embeddings

- Universal Sentence Encoder (USE)
  - USE is utilized to build our **baseline model**
  - State-of-the-art feature that ranked first at HatEval campaign [Indurthi et al., 2019]
  - Each tweet is represented by a vector size of 512



Methodology: our MWE identification

We build an **MWE lexicon** by extracting MWE from STREULSE corpus [Schneider and Smith, 2015][3]

- Obtained MWE lexicon contains **1855** MWEs classified into **20** categories

This MWE lexicon is used to tag MWEs in tweets:

- *shut up* –> Full verb-particle construction
- *tax payer* –> Nominal
- *stand for* –> Inherently adpositional verb
- *give a crap* –> Verbal idiom
- *thank you* –> Adverb

We add a **special category** for words that are not belonging to an MWE

[3] A corpus and model integrating multiword expressions and supersenses

Methodology: MWE representation

MWE features are represented by lexical categories and words embeddings:

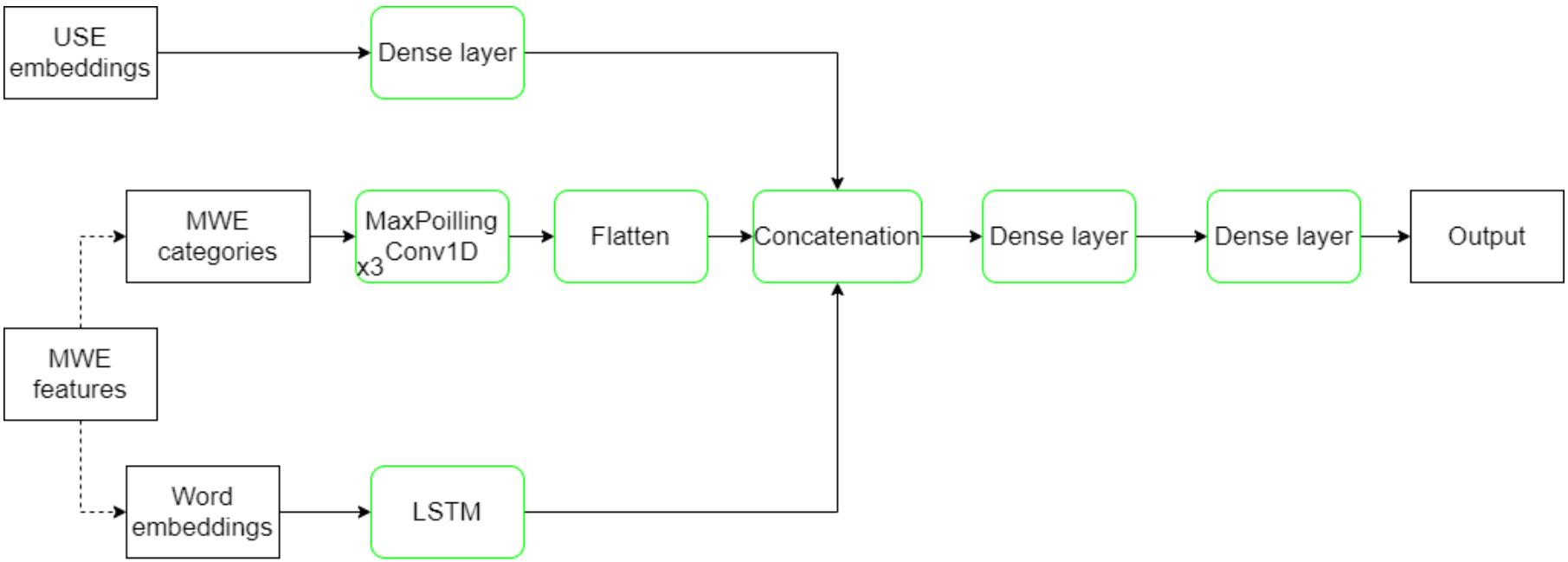
- We associate to each word of the tweet the MWE category (one-hot encoding)

Lets	***get***	***out***	*of*	*EU*	*NHS*	*Schools*
O	*VPC.full*	*VPC.full*	*O*	*O*	*O*	*O*

- We associate a word embedding to each word composing an MWE
  - We experiment static and dynamic word embeddings (WE)
    - Word2vec as static embeddings
    - BERT as dynamic embeddings

$$WE(\textbf{get}) \quad WE(\textbf{out})$$

Methodology: proposed model



Experimental setup: HatEval corpus

SemEval2019 task 5: HatEval English corpus [Basile et al., 2019][4]

- **Binary classification**: **hate** and **non-hate**
- 13k tweets divided in train (9k), dev (1k) and test (3k) set
- Each set contains **42%** of hateful tweets.
- Pre-processing:
  - We keep the case unchanged
  - We remove hashtags, mentions and URLs
- We use a small part of the training set for validation

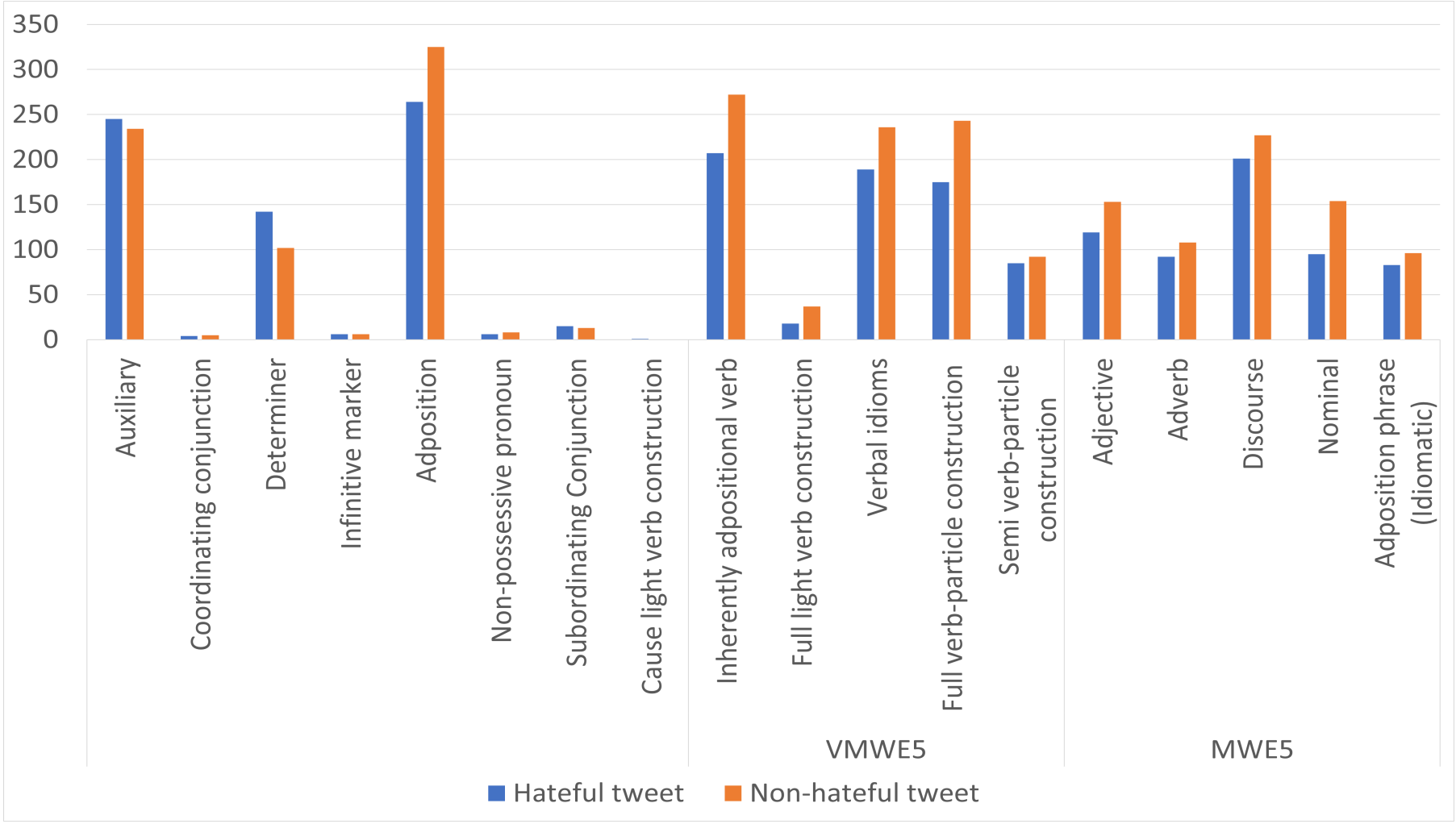
[4] SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter.

Experimental setup: MWE statistic

	MWE categories	Hateful	Non-hateful	Both
MWE5	Adjective	9	8	255
	Adverb	1	5	194
	Discourse	12	15	401
	Nominal	25	36	189
	Adposition phrase (idiomatic)	9	36	134
VMWE5	Inherently adpositional verb	11	21	447
	Full light verb construction	9	10	36
	Verbal idioms	14	24	384
	Full verb-particle construction	11	20	387
	Semi verb-particle construction	6	18	153
	Auxiliary	4	0	475
	Coordinating conjunction	1	0	8
	Determiner	1	2	242
	Infinitive marker	0	0	12
	Adposition	3	13	573
	Non-possessive pronoun	0	3	11
	Subordinating conjunction	0	0	28
	Cause light verb construction	1	0	0

Experimental setup: MWE statistic

25% of tweets contain at least one MWE



Experimental setup: evaluation metric and training setup

- Evaluation metric:
  - We use macro-F1 score which is the average of the F1 scores of all classes
- Training setup:
  - We train 9 models for each experiement
  - We take the best model on dev set
  - Then, we predict on the test set

Results: HatEval test set

The best system (FERMI) at the HatEval campaign obtained 65% of macro-F1 [Indurthi et al., 2019][5]

Features	**HatEval**		
	F1		**Macro-F1**
	Hateful	Non-hate	
USE	64.9	66.4	65.7
USE, MWEall, word2vec	64.5	68.2	66.3
USE, VMWE5, MWE5, word2vec	66.1	67.0	66.5
USE, MWEall, BERT	64.2	69.4	**66.8**
USE, VMWE5, MWE5, BERT	64.8	68.2	66.5

–> MWE features based system outperforms the baseline

[5]FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter.

Results: Founta corpus

- Founta corpus contains 100K tweets annotated with normal, abusive, hateful and spam labels
- We focus on **normal**, **abusive** and **hateful** tweets (86k tweets)
- We split corpus in 60% train, 20% validation and 20% test sets
- Each set contains about 62%, 31%, and 6% of normal, abusive, and hateful tweets
- We apply the same pre-processing as for the HatEval corpus

Results: Founta test set

Features	Founta			
	F1			Macro-F1
	Norm	Abus	Hate	
USE	94.2	87.8	34.6	72.2
USE, MWEall, word2vec	93.8	86.9	36.5	72.4
USE, VMWE5, MWE5, word2vec	93.9	87.1	37.2	72.7
USE, MWEall, BERT	94.0	87.1	37.5	72.9
USE, VMWE5, MWE5, BERT	93.8	86.9	38.2	**73.0**

–> These results are consistent with those observed on the Hat-Eval test set

Results: tweets containing at least one MWE

Features	HatEval		
	F1		Macro-F1
	Hateful	Non-hate	
USE	67.8	62.3	65.0
USE, MWEall, word2vec	71.7	61.4	66.6
USE, MWEall, BERT	73.9	61.3	**67.6**

Features	Founta			
	F1			Macro-F1
	Normal	Abusive	Hateful	
USE	91.1	94.1	41.6	75.6
USE, MWEall, word2vec	91.4	86.9	44.6	**76.5**
USE, MWEall, BERT	90.9	94	43.3	76.1

–> These results show that MWE features allow to enrich our baseline system

Conclusion

- We studied the impact of new features for hate speech detection: **MWE**
- Our proposed approach outperforms USE-based FERMI[6] system
- The best configuration using MWE features is MWEall with BERT embeddings

[6] Ranked first at HatEval campaign

References

V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/S19-2007`.

V. Indurthi, B. Syed, M. Shrivastava, N. Chakravartula, M. Gupta, and V. Varma. FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 70–74, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/S19-2009`.

J. T. Nockeby. Hate speech. In L. W. Levy, K. L. Karst, and D. J. Mahoney, editors, *Encyclopedia of the American Constitution*, pages 1277–1279. Macmillan 2nd edition, 2000.

N. Schneider and N. A. Smith. A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1537–1547. ACL, 2015.