THE PROBLEM OF DETERMINING AGE FROM TWEETS

The challenge: To build a model to determine the age group of the tweeter user using the information in a tweet (13-17,17-24,24-35,35-45,45-65,65-XX years)

- Different roles in the society determine the language use [1]
- Age are shaped depending on the societal context [2]
- On twitter depending of the context users may emphasize specific aspects which leads to linguistic variation [3]

[1] (Eckert, 2008)

[2] (Bucholtz and Hall, 2005)

[3] (Nguyen,2014)

The problem of determining Age from Tweets

The challenge: To build a model to determine the age group of the tweeter user using the information in a tweet (13-17,17-24,24-35,35-45,45-65,65-XX years)

- Different roles in the society determine the language use [1]
- Age are shaped depending on the societal context [2]
- On twitter depending of the context users may emphasize specific aspects which leads to linguistic variation [3]

[1] (Eckert, 2008)

[2] (Bucholtz and Hall, 2005)

[3] (Nguyen,2014)

The problem of determining Age from Tweets

The challenge: To build a model to determine the age group of the tweeter user using the information in a tweet (13-17,17-24,24-35,35-45,45-65,65-XX years)

- Different roles in the society determine the language use [1]
- Age are shaped depending on the societal context [2]
- On twitter depending of the context users may emphasize specific aspects which leads to linguistic variation [3]

[1] (Eckert, 2008)

[2] (Bucholtz and Hall, 2005)

[3] (Nguyen,2014)

The problem of determining Age from Tweets

The challenge: To build a model to determine the age group of the tweeter user using the information in a tweet (13-17,17-24,24-35,35-45,45-65,65-XX years)

- Different roles in the society determine the language use [1]
- Age are shaped depending on the societal context [2]
- On twitter depending of the context users may emphasize specific aspects which leads to linguistic variation [3]

[1] (Eckert, 2008)

[2] (Bucholtz and Hall, 2005)

[3] (Nguyen,2014)

Proposal

Three Approaches were tested:

- A quantification of the tweet elements (CtTweets)
- A Convolutional neural network trained from scracth (CNN)
- A Transfer learning model using the BERT model (BERT)

Proposal

Three Approaches were tested:

- A quantification of the tweet elements (CtTweets)
- A Convolutional neural network trained from scracth (CNN)
- A Transfer learning model using the BERT model (BERT)

F(x) = [#hashtags, #words, #users, #upper letters,# low letters,# symbols,bool(url), tweet length, length short word, length large word]

Proposal

Three Approaches were tested:

- A quantification of the tweet elements (CtTweets)
- A Convolutional neural network trained from scracth (CNN)
- A Transfer learning model using the BERT model (BERT)

Proposal

Three Approaches were tested:

- A quantification of the tweet elements (CtTweets)
- A Convolutional neural network trained from scracth (CNN)
- A Transfer learning model using the BERT model (BERT)

embedding,normalization,convolutional layer, max pooling,average,dense layer with softmax

PROPOSAL

Three Approaches were tested:

- A quantification of the tweet elements (CtTweets)
- A Convolutional neural network trained from scracth (CNN)
- A Transfer learning model using the BERT model (BERT)

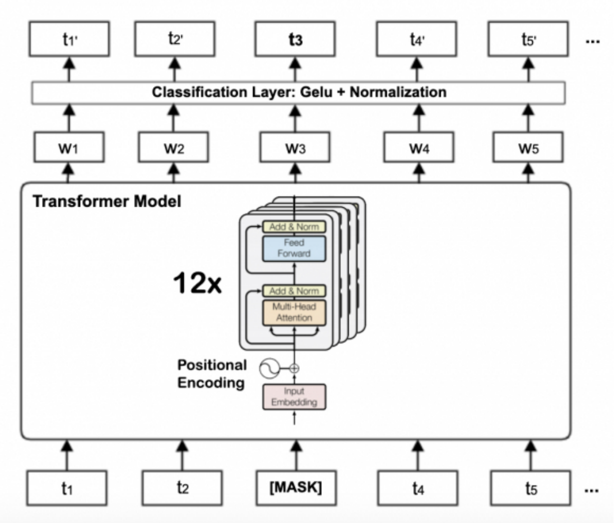
Proposal

Three Approaches were tested:

- A quantification of the tweet elements (CtTweets)
- A Convolutional neural network trained from scracth (CNN)
- A Transfer learning model using the BERT model (BERT)

Using the english vocab weights in a BERT model

The Bert Model



Experiments

- The data is separated in Train (65%), Validation(10.5%) and Test (24.5%)
- The data is clean, hashtags are word separated, symbols removed, lower case is used, stop words and URLs eliminated, words are lemmatized (18 min)
- A data augmentation process was performed in the CtTweets model

Experiments

- The data is separated in Train (65%), Validation(10.5%) and Test (24.5%)
- The data is clean, hashtags are word separated, symbols removed, lower case is used, stop words and URLs eliminated, words are lemmatized (18 min)
- A data augmentation process was performed in the CtTweets model

Experiments

- The data is separated in Train (65%), Validation(10.5%) and Test (24.5%)
- The data is clean, hashtags are word separated, symbols removed, lower case is used, stop words and URLs eliminated, words are lemmatized (18 min)
- A data augmentation process was performed in the CtTweets model

Results

	Multiclass Problem		**Binary Problem**	
	Accuracy	**F1-Score**	**Accuracy**	**F1-Score**
CtTweet (3.7seg)	0.380	0.215	0.695	0.691
CtTweet (Data-augmented)	0.351	0.251	0.695	0.690
CNN (7.3seg)	0.371	**0.284**	0.700	0.71
BERT (31min)	**0.420**	0.243	**0.765**	**0.798**

Conclusions

- Results using a single tweet are still weak. The binary group is promising
- Future improvements include translating emojis to words
- Training a model per group could improve the results
- Build a supervised vocabulary would improve current approaches also as more variables

Thanks...