Is global warming real ?
Machine Learning for Natural Language Processing 2020

YOUSSFI Salim
ENSAE Paris
salim.youssfi@ensae.fr

BEUCHER Tristan
ENSAE Paris
tristan.beucher@ensae.fr

Access to the Collab Notebook : https://colab.research.google.com/drive/14csUH6S65KrxBHVYUxzjvSn90AdpHfzu

Access to the repo Github : https://github.com/salimYOU/Projet_NLP

1 Purpose of the study and presentation of the data

In general, global warming is an established fact accepted by the majority of the population and governments. Nevertheless, there is still a part of the population that refuses to believe in this theory..

The main objective of this study is to study the arguments of the so-called 'climate sceptics'. For this purpose, we have at our disposal a database of tweets about climate change. Some of these tweets deny the existence of climate change, others do not. Each tweet is labelled as follows: 'yes' if it suggests that global warming is real; 'no' if he's suggesting global warming is does not exist (or is not related to human activity); 'I can't tell' if the tweet is ambiguous.

Of the 5539 usable tweets, 2821 tweets come from people who believe in climate change and 1035 tweets come from climate skeptics. The rest of the tweets (1683) are considered neutral.

Our overall approach is to find a statistical model of these tweets to derive an argument specific to climate sceptics. It can be the subjects mentioned by these people, the buzz words, or the tone of their speeches.

2 Tweets representation

In a first step, we tried to get a 'coherent' representation of the tweets. Our approach is divided into two distinct steps: the tokenization of tweets; the vector representation of tweets.

The tokenization step is mainly about cleaning each tweet to represent it as a list of tokens. It is mainly about eliminating the parasitic symbols, the punctuation of tweets, as well as performing a normalization of the words via the 'wordnet' module. A lemmatization operation has also been performed to associate words that we want to consider as peers (like 'global warming').

After tokenization and cleaning of the tweets, we represented the tweets in different ways. The idea is to represent each tweet by a numerical vector that should allow us to extract a maximum of information. We tested five representations of a tweet. The models *TF-IDF* and *Word2Vec* are trained on the database of tweets. Since the number of tweets (5000) is relatively low, we also used pre-trained models such as *Fast2Vec* and *BERT*.

In reality, each model returns a numerical vector for a single word, not for an entire tweet. For the models *TF-IDF*, *Word2Vec* and *Fast2Vec*, we have, for example averaged the representations of the words obtained in each tweet.

3 Tweets classification

How to evaluate these different representations? Let's not forget that our final objective is to study the arguments of climate sceptics. We therefore decided to evaluate the different representations in relation to their ability to recognize a 'No' (climate sceptic) tweet. To do this, we trained classification models based on the representations of the five methods described above. Each time, the training was performed on a fraction of the base (66%) and tested on the remaining fraction (34%).

We considered the classic RandomForest, SVM and XGBoost models. For the reason explained above, the metric we used is the F1-score on the 'No' class. The 5 best results (out of the possible 3×5) are :

'model' and 'representation'	F1-score
XGBoost and BERT	0.53
RandomForest and Fast2VecCluster	0.51
XGBoost and Word2Vec	0.49
XGBoost and Fast2VecAverage	0.46
RandomForest and Word2Vec	0.46

To assess whether these scores are 'good', we compared them to the result of the LSTM (Long Short Term Memory) model, which makes its own representation of tweets. This model has an F1-score of 0.58 on the 'No' classification. We can therefore consider that our representations are quite good (knowing that we did not perform a cross validation on the XGBoost, SVM and RandomForest models).

4 Tweets analysis

It is now a matter of using the previous tools to analyze in detail the content of tweets.

A first descriptive analysis of the tweets can be done after the tokenization step (even before the different representations). For example, we can look at the frequency of words in the tweets. We notice that words associated with media scandals and lies are much more present in the tweets of climato-skeptics. For example, the word 'hoax' appears in 2.71% of tweets rated 'No' but only in 0.21% of tweets rated 'Yes'. There are also many scandals such as the 'climate_gate' (Climatic Research Unit email controversy) which refers to email hacking at the University of East Anglia. These emails would have shown that global warming is a scientific conspiracy.

Concerning the tweets that believe in climate change, we can globally observe the theme related to the effort (could, energy, fight) of solidarity (us, help, world) and the environment (green, earth, carbon).

To study in more detail the association of words between them, we considered two approaches: a first one consisting in clsuterizing tweets from the different representations; a second approach consisting in using the LDA (Latent Dirichlet allocation) model to bring up subjects.

The clustering of tweets was done on the representation that seemed most relevant to us (see section 3). We therefore retained the word-embedding done by the BERT model. We used the K-means aglrithm, setting the number of clusters to three so as not to scatter too much. Two clusters are well interpretable on the database of climato-skeptics. A first cluster clearly highlights the theme of fraud and conspiracy (we find in particular the words 'fraud', 'hoax' or 'theory'). Another cluster refers more to the political world with the over-representation of words like 'obama', 'al_gore' (former Vice President of the USA). These words are associated with negatives terms such as 'silly' or 'scam'..

The application of the LDA model (done on three topics) also brings out a topic related to lying and hoaxes. The terms related to scandals and politicians are less present in the output of this model.

We are also looking for information on the tone of the tweets. For this, we used the SentiWordNet scoring model which scores a word according to its 'positive', 'negative' or 'neutral' character. According to this model, climate sceptics are more demonstrative. They use stronger and more striking terms than other people.

5 Conclusion

In conclusion, our study showed that climate sceptics have a line of argument based mainly on the controversies and scandals related to climate change. According to them, all this seems to be a big lie. The political class is the main guarantor of this. Finally, this argumentation seems to be marked by 'shock phrases' and strong positions.