Can sentiment and the writing structure information from French social media conversations be retrieved?

Sije van der Veen

***Abstract*—**

I. INTRODUCTION

The idea for this project originated by Radio Netherlands worldwide media (RNW). RNW creates communities online to contribute to social change. The topics they cover reflect young peoples needs in relation to love, relationships and their hopes and ambitions for their societies. Local teams of media-makers manage the digital communities. In turn, local teams build and coordinate networks of young people who produce content to engage the wider community across our thematic areas [**?**].

This research can contribute to retrieve more and in depth information from the content on the discussion platforms. In this research will be worked with French data, the current data method which is used to store and explore the data already provides information, like sentimental and emotional information. This information is based on predictions, by observing the data manually it is easy to observe that the model. Therefore in this study a new models will be developed to optimize the results for sentiment labeling. is applied on the data. Besides a better predicting method for sentiment information, also topic modeling is applied on the data. By applying topic models on these discussion data, it would be interesting to see if topic modeling is a suitable methods to distinguish between comments, how likely they are with the post. This information could help to see if the post are written same writing style of their audience.

For this study the research question is defined as followed: Can sentiment and the writing structure information from French social media conversations be retrieved. To split the question in pieces, the following sub questions are defined:

1) Can the sentiment be predicted for French Twitter data?
2) Will probabilities for each label per word optimize the results?
3) Will using a Recurrent Neural network optimize the results?
4) Can Facebook conversations be compared, using the the likeliness of the post as gold standard?
5) Is topic modeling a good measurement for likeness of Facebook posts and comments?

II. RELATED WORK

A. Stems and Phrases

Before a text data can be used for language model, cleaning methods can help to optimize the performance. Words can have different morphological variations as in inflectional (plurals, tenses) or derivational (making words nouns). In general they have the same meaning or almost the same meaning. Stemmers can be used to reduce the morphological variations of similar words to a common stem. In many causes suffixes will be removed. The effect of stemming is general small but significant, which can be in some languages crucial. There are two ways to perform stemming; by using a list of related words (dictionary) or using an program which determines words (algorithm). Algorithms can make easily mistakes, for example with by removing and assuming it is plural. In cause of supplies to supplie is a false negative and ups to up results in a false positive.

Phrases are more precise than single words. For example documents containing black sea vs. two words black and sea. And ambiguous for example big apple vs apple. Phrases can be recognized in two different approaches, by identifing syntactic phrases using a part-speech (POS) tagger or by using word n-grams. a simple data-driven approach, where phrases are formed based on the unigram and bigram counts. Bigrams with scores above the chosen threshold are then used as phrases. POS taggers use statistical models of text to predict syntactic tags of words These tags can be used to find phrases in textual data. POS tagging is too slow for large collections. Frequent n-grams are more likely to be meaningful phrases.

B. Language model fundamentals

Language models are used on many languages, mainly frequently spoken languages, and for many purposes. To estimate the relative likelihood of different phrases is useful in many natural language processing models. Languages models are used for speech recognition, part-of-speech-tagging, handwriting recognition and many other applications. Bag of words representation/ unigram are commonly used for query likelihood models. A separate language model is associated with all documents in a collection. The documents are ranked with the probability of a given query in a language model. A major problem in building a language model is the sparsity of data. In the training of the model not all words are observed, which results in zero probabilities. This can be solved by smoothing techniques, assuming that a word is depending on the previous word(s) in a sentence (n-gram model). Words in bigram and trigram denotes for a language model with n = 2 and n = 3.

Different type of models were used for estimating continuous representation of words. Two well known models are Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA).