**Vectorization**

Artificial Intelligence and the public opinion

Maya Mileva

HW2 IST 736

A circuit board

Description automatically generated

1. **Introduction**

It is an era where information is just a click away. People are influenced by it and want to share thoughts about it. This is where social media enters. Social media is a part of people`s day-to-day life that can’t be ignored. It has both positive and negative effects. Most of the people are addicted to many social media platforms like Facebook, Twitter, Instagram, etc. and it is considered odd not to be connected. Social media has grown to become a large platform for entrepreneurs, businesses, organizations and various other professionals who seek identification and recognition at a moderate cost.

It’s estimated that 80% of the world’s data is unstructured, in other words it’s unorganized. Huge amounts of text data (emails, support tickets, chats, social media conversations, surveys, articles, documents), is created every day but it’s hard to analyze, understand, and sort through, not to mention time-consuming and expensive. Sentiment analysis, however, helps businesses make sense of all this unstructured text by automatically tagging it. [Sentiment](https://www.monkeylearn.com/sentiment-analysis-online) analysis  is the interpretation and classification of emotions (positive, negative and neutral) within text data using text analysis techniques and allows businesses to identify customer sentiment toward products, brands or services in online conversations and feedback. The word sentiment means a view of or attitude toward a situation or event, an opinion. For example, different forms of music convey different kinds of sentiment.

From SIRI to self-driving cars, artificial intelligence (AI) is progressing rapidly and is one of the most popular topics in social media today. AI is an area of computer science that emphasizes the creation of intelligent machines that work and react like humans. Some of the activities computers with artificial intelligence are designed for include speech recognition, learning, planning, problem solving. In other words, AI is using  computers and machines to do things humans can’t do as well like process huge amount of data. Though people love technology so deeply, they sure have an anxious relationship with it. They are worried about AI stealing their job, their house, or their significant other. There is a fundamental master-servant relationship at play here. Not only does AI depend on people to create it, but it also needs them to provide it with data and give it tools for interpreting and interacting with that data. Artificial intelligence systems need people to direct them before they can provide any sort of value at all.

1. **Analysis and Models**

**2.1 About the data**

Four data sets were used. The first one was a small document, consisting of text files, created only for the purpose of showing basic aspects of sentiment preprocessing like tokenization, frequency distribution, stop words removal and stemming. The second one text was created to imitate positive and negative tweets about Artificial Intelligence, containing two labeled  mini corpuses of five text files each. Third corpus was added from positive and negative movie reviews for the purpose of evaluation of the accuracy of the different Sentiment Analysis Tools. The fourth one was obtained from <http://www.cs.cornell.edu/people/pabo/movie-review-data/> and consist of corpuses of 700 positive and 700 negative reviews. The data set was added also for comparison and evaluation purposes, while is slightly bigger than the rest. The data was labeled and combined to form a data frame of  1400 labeled reviews.

There are many ways to examine the context of a text apart from simply reading it. Text 1 and Text. 2 were created with simple sentences, and are about dog, the affection toward the animal and walks.

**Text1.**

|  |  |
| --- | --- |
| ***I love my dog. My dog is my best friend, he is awesome. Me and the dog go for a walk every day and my dog meet all his friends. I love my dog`s blue eyes and that he is so good with people. I love him so much.*** | A picture containing drawing  Description automatically generated |

Word clouds  shows the most frequent words in text, by displaying them bigger than the rest.

A close up of a logo

Description automatically generated

******Text 2.

***I love walking outside. It is really nice and refreshing and I burn calories. I walk at least 3 miles every day with my dog. We enjoy the nice weather in Florida during out winter walks.***

Word clouds  shows the most frequent words in text, by displaying them bigger than the rest.

A screenshot of a cell phone

Description automatically generated

Word cloud for positive and negative tweets are displayed next. It is interesting way to visualize what people talk about most.

Negative Tweets

A close up of a piece of paper

Description automatically generated

Positive Tweets

A close up of a map

Description automatically generated

* 1. **Data Transformation**

When text is mentioned, few aspects have to be considered. At one level, text is a sequence of symbols on a page such as this one. At another level, it is a sequence of chapters, made up of a sequence of sections, where each section is a sequence of paragraphs, and so on. For sentiment analysis text is considered as nothing more than a sequence of words and punctuation.

In order to be processed and analyzed, those words have to become tokens. Tokenization is the process by which big quantity of text is divided into smaller parts.

The goal of the project is to identify the public sentiment towards AI on social media. Decision about what to count and how to count have to be made. There are different vectorization options. The CountVectorizer provides a simple way to both tokenize a collection of text documents and build a vocabulary of known words, but also to encode new documents using that vocabulary.

**text = ["I love my dog. My dog is my best friend, he is awesome. Me and the dog go for a walk every day and my dog meet all his friends. I love my dog`s blue eyes and that he is so good with people. I love him so much."]**

**tokens** ['I', 'love', 'my', 'dog', '.', 'My', 'dog', 'is', 'my', 'best', 'friend', ',', 'he', 'is', 'awesome', '.', 'Me', 'and', 'the', 'dog', 'go', 'for', 'a', 'walk', 'every', 'day', 'and', 'my', 'dog', 'meet', 'all', 'his', 'friends', '.', 'I', 'love', 'my', 'dog', '`', 's', 'blue', 'eyes', 'and', 'that', 'he', 'is', 'so', 'good', 'with', 'people', '.', 'I', 'love', 'him', 'so', 'much', '.']

**My vocabulary is…**

**{'love': 18, 'my': 22, 'dog': 6, 'is': 17, 'best': 3, 'friend': 10, 'he': 14, 'awesome': 2, 'me': 19, 'and': 1, 'the': 26, 'go': 12, 'for': 9, 'walk': 27, 'every': 7, 'day': 5, 'meet': 20, 'all': 0, 'his': 16, 'friends': 11, 'blue': 4, 'eyes': 8, 'that': 25, 'so': 24, 'good': 13, 'with': 28, 'people': 23, 'him': 15, 'much': 21}**

Using CountVectorizer is noticeable that all words were made lowercase by default and that the punctuation was ignored. The encoded vector is a sparse matrix. An array version of the encoded vector is showing a count of word occurrences.

[[1 3 1 1 1 1 5 1 1 1 1 1 1 1 2 1 1 3 3 1 1 1 5 1 2 1 1 1 1]]

The output of word tokenization can be converted to Data Frame for better text understanding in machine learning applications. It can also be provided as input for further text cleaning steps such as punctuation removal, numeric character removal or stemming. The encoded vectors from CountVectorizer can then be used directly with a machine learning algorithm. Machine learning models need numeric data to be trained and make a prediction. Word tokenization is crucial part of the text (string) to numeric data conversion.

When analyzing text, often become important to know how words of a text that are most informative about the topic and genre of the text can be automatically identified. Frequency distribution is one of the answers: shows the frequency of each vocabulary item in the text. In the first text example, frequency distribution plot looks like Fig.1. Word dog is used the most, after that is ‘.’ Punctuation marks are still tokens. Depending on the purpose of the analysis, often punctuation is removed.

**Fig. 1**A close up of text on a white background

Description automatically generated

An alternative to calculate word frequencies is called **TF-IDF**. This is an acronym than stands for “*Term Frequency – Inverse Document*” Frequency which are the components of the resulting scores assigned to each word.

* **Term Frequency**: This summarizes how often a given word appears within a document.
* **Inverse Document Frequency**: This downscales words that appear a lot across documents.

TF-IDF are word frequency scores that try to highlight words that are more interesting, e.g. frequent in a document but not across documents.

**text = ["I love my dog. My dog is my best friend, he is awesome. Me and the dog go for a walk every day and my dog meet all his friends. I love my dog`s blue eyes and that he is so good with people. I love him so much.",**

**"I love walking outside. It is really nice and refreshing and I burn calories. I walk at least 3 miles every day with my dog. We enjoy the nice weather in Florida during out winter walks."]**

**[[0.1226168 0.26172859 0. 0.1226168 0.1226168 0.1226168 0. 0. 0.08724286 0.43621432 0. 0. 0.08724286 0.1226168 0. ….]]**

The scores are normalized to values between 0 and 1 and the encoded document vectors can then be used directly with most machine learning algorithms.

Another decision is about stop words. A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. These words are taking up space in the working database or are taking up valuable processing time. For this, they can be removed by storing a list of words that is considered to be stop words. Stop words list can be altered and modified, depending on the corpus and the final goal of each analysis.

In order to get better pictures with the most frequent words, stop words have been removed and is displaying the new frequency distribution plot.

**Fig. 2**A screenshot of a cell phone

Description automatically generated

After removing stop words and punctuations the results are slightly more interesting. It is clear now that the corpus id about love, dog and best friend.

When the corpus is bigger another approach is to look at Bigram Frequencies and Mutual Information. People read texts. The texts consist of sentences and also sentences consist of words. Human beings can understand linguistic structures and their meanings easily, but machines are not successful enough on natural language comprehension yet. So, people try to teach some languages to machines like they do for an elementary school kid. This is the main concept; words are basic, meaningful elements with the ability to represent a different meaning when they are in a sentence. Sometimes word groups provide more benefits than only one word when explaining the meaning.

Another approach is stemming and lemmatization. They are both ways to reduce  words down to the base word. Stemming removed redundancy in the data and variations in the same word. Lemmatization transform the root to a real word.

After some basic sentiment analysis was discussed and performed , is time to incorporate tweeter examples from the second corpus. When stemming the tweets, no important linguistic information was removed. For best vectorization option CountVectorized was considered.

**NLTK sentiment analysis:**

Classifier based on the Naive Bayes algorithm, what it does is simply  classify texts or parts of texts into a pre-defined **sentiment**.  Sentiment analysis is becoming a popular area of research and social media analysis, especially around user reviews and tweets. The corpuses discussed in this paper are really small, each positive and negative review or tweet contain only 5 texts, which make separation for training and testing set really difficult. Due to the small samples the result won`t be very objective. The fourth new corpus was added to help that issue.

**Tweets**:

|  |
| --- |
| *Training classifier*  *Evaluating NaiveBayesClassifier results...*  *Accuracy: 0.5*  *F-measure [neg]: 0.6666666666666666*  *F-measure [pos]: None*  *Precision [neg]: 0.5*  *Precision [pos]: None*  *Recall [neg]: 1.0*  *Recall [pos]: 0.0* |

**Movie Reviews:**

|  |
| --- |
| *Training classifier*  *Evaluating NaiveBayesClassifier results...*  *Accuracy: 0.5*  *F-measure [neg]: 0.6666666666666666*  *F-measure [pos]: None*  *Precision [neg]: 0.5*  *Precision [pos]: None*  *Recall [neg]: 1.0*  *Recall [pos]: 0.0* |

The classifier did not do very good job with the small data set. The reason is not sufficient data for prediction, due to the small size is harder to split the train and test set and at the same time get meaningful results.

New data set was added.

**Movie Reviews – BIG data:**

|  |
| --- |
| *Training classifier*  *Evaluating NaiveBayesClassifier results...*  *Accuracy: 0.7674897119341564*  *F-measure [neg]: 0.7797270955165693*  *F-measure [pos]: 0.7538126361655773*  *Precision [neg]: 0.7380073800738007*  *Precision [pos]: 0.8046511627906977*  *Recall [neg]: 0.8264462809917356*  *Recall [pos]: 0.7090163934426229* |

The classifier did better job here, the accuracy is 77%. Obviously larger data set have to be considered using it. The precision and recall are determined, so the right classified reviews are calculated. A confusion matrix can be created to clearly show the classification.

Two rule-based methods for Sentiment Analyses will be discussed:

* **VADER**: Parsimonious rule-based model for sentiment analysis of social media text.
* **Text Blob**: Simple rule-based API for sentiment analysis

**VADER:**

Two different data set were used to test the accuracy.

VADER belongs to a type of sentiment analysis that is based on lexicons of sentiment-related words. In this approach, each of the words in the lexicon is rated as to whether it is positive or negative, and in many cases, how positive or negative. Below is an excerpt from VADER’s lexicon from tweets and movie reviews corpuses. More positive words have higher positive ratings and more negative words have lower negative ratings.

**Tweets**:

|  |  |  |
| --- | --- | --- |
| label | compound | sentence |
| pos | 0.8481 | I trust Artificial Intelligence with handling medication … |
| pos | 0.8221 | Artificial Intelligence if the future, I really love it… |
| pos | 0.9463 | Siri is my best friend, she never lies to me, buy groceries… |
| pos | 0.6435 | So excited and happy about Artificial Intelligence, the future… |
| pos | 0.8316 | Artificial Intelligence can significantly augment the decision…. |
|  |  |  |
| neg | -0.8516 | I hate hate hate Artificial intelligence!.. |
| neg | -0.8297 | I can never trust Artificial Intelligence to tell me … |
| neg | -0.4215 | I will lose my job because of nasty Artificial Intelligence… |
| neg | 0.2462 | Artificial intelligence will crash the economy!!.. |
| neg | 0.3382 | Artificial intelligence is ruining my life, taking my job away!.. |

**Movie Reviews:**

|  |  |  |
| --- | --- | --- |
| label | compound | sentence |
| pos | -0.5887 | it's surprising to see how much more it looks like a … |
| pos | 0.8825 | did anybody know this film existed a week before it opened?... |
| pos | 0.9964 | in order to make the film a success , all they had to do was… |
| pos | 0.9868 | the movie opens with blackness , and only distant… |
| pos | -0.3525 | Jackie doesn't even have enough money for a haircut , looks like… |
|  |  |  |
| neg | -0.8481 | even the magic kingdom at its most mediocre -- that'd be … |
| neg | 0.6824 | unsuccessfully attempting to gain the woman's favor… |
| neg | -0.9879 | john carpenter apparently believes that action scenes… |
| neg | -0.9753 | is it about a wholesome surveillance man who loses sight… |
| neg | 0.7836 | it's a terrible mess of a movie starring a terrible… |

VADER doesn’t just do simple matching between the words in the text and in its lexicon. It also considers certain things about the way the words are written as well as their context. In the example about, better results were completed with the smaller corpus with tweets. One of the things that VADER recognizes is capitalization, which increases the intensity of both positive and negative words and also considers what happens when modifying words are present in front of a sentiment term. VADER isn’t the most accurate at classification, but it is a useful tool for  creating labels out of unlabeled data.

**TextBlob**

TextBlob is a popular Python library for processing textual data. It is built on top of NLTK[,](http://www.nltk.org/) another popular Natural Language Processing toolbox for Python. TextBlob uses a sentiment lexicon (consisting of predefined words) to assign scores for each word, which are then averaged out using a weighted average to give an overall sentence sentiment score. Three scores: **“polarity”**, **“subjectivity”** and “intensity” are calculated for each word. Some intuitive rules are hardcoded inside TextBlob to detect modifiers (such as adverbs in English: “*very good*”) that increase or decrease the overall polarity score of the sentence.

**Tweets:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| label | prediction | sentiment | len | excerpt | tags |
| neg | neg | -0.787500 | 41 | I hate hate hate Artificial. | [(I, PRP), (hate, VBP), ... |
| neg | neg | -0.433333 | 105 | I can trust Artificial... | [(I, PRP), (can, MD), … |
| neg | neg | -0.558974 | 61 | I will lose my job because... | [(I, PRP), (will, MD), … |
| neg | neg | -0.937500 | 48 | Artificial intelligence will … | [(Artificial, JJ)... |
| neg | neg | -0.750000 | 63 | Artificial intelligence is … | [(Artificial, JJ), … |
|  |  |  |  |  |  |
| pos | neg | -0.600000 | 150 | I trust Artificial Intelligence. | [(I, PRP), (trust, VBP), (Ar... |
| pos | neg | -0.033333 | 83 | Artificial Intelligence if ... | [(Artificial, JJ), (Intellig... |
| pos | pos | 0.300000 | 132 | Siri is my best friend, she... | [(Siri, NNP), (is, VBZ), ... |
| pos | pos | 0.066667 | 72 | So excited and happy ... | [(So, RB), (exited, JJ), ... |
| pos | pos | 0.135000 | 275 | Artificial Intelligence can... | [(Artificial, JJ), (... |

**Movie Reviews:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| label | prediction | sentiment | len | excerpt | tags |
| neg | pos | 0.025467 | 2929 | " quest for camelot " is... | [(quest, JJS), (for, IN)... |
| neg | pos | 0.022925 | 4418 | synopsis : a mentally... | [(synopsis, NN), (a, DT)… |
| neg | pos | 0.043234 | 3911 | capsule : in 2176 on the... | [(capsule, NN), (in, IN)… |
| neg | pos | 0.003334 | 3365 | so, ask yourself what " ... | [(so, RB), (ask, VB), , ... |
| neg | neg | -0.054577 | 3554 | that's exactly how long t .... | [(that, DT), ('s, VBZ)... |
|  |  |  |  |  |  |
| pos | pos | 0.023663 | 4227 | films adapted from comic... | [(films, NNS), (adapted... |
| pos | pos | 0.103847 | 4096 | every now and then a = ... | [(every, DT), (now, RB)… |
| pos | pos | 0.131092 | 2421 | you've got mail works a ... | [(you, PRP), ('ve, VBP)... |
| pos | pos | 0.110626 | 6092 | " jaws " is a rare film that... | [(jaws, NN), (is, VBZ), (a… |
| pos | neg | -0.070151 | 3898 | moviemaking is a lot like... | [(moviemaking, NN), (is, ... |

In the example above Texblob didn`t do very good job predicting the sentiment, especially with the negative movie reviews. All negative tweets were correctly predicted, which is really interesting. It seems that Textblob predict 90% positive movie review. There are different approaches and different features in predicting tweets and movies and that might be the problem here.

1. **Results**

The results weren`t convincing enough for even evaluating those three models. . NLTK with Naïve Bayes showed better results classifying the big data set, than relatively small ones. The Tweets and small Movie Reviews datasets were really small to lead to any conclusions. VADER did slightly better job, but accurate prediction is impossible. For cases with small data sets, or when the sentiment for couple of sentences have to be predicted, tools like [**SentiStrength**](http://sentistrength.wlv.ac.uk/results.php?text=I+love+you+but+hate+the+current+political+climate.&submit=Detect+Sentiment&result=dual)and [**Sentiment Analysis with Python NLTK Text Classification**](http://text-processing.com/demo/sentiment/)can be very useful. For accurate models and prediction more data have to be collected in order to have meaningful training sample for the classifier. The data can be from tweets, Facebook comments, Instagram tags. The buzz about AI is everywhere, people are enjoying social media, because they can be finally heard. All preprocessing steps are the similar for bigger corpuses – tokenization, stemming, feature creations. CountVectorizer was introduced to  convert the collection of text documents to a matrix of token counts. The goal of the analyses determines those preprocessing steps and the choice of the tools used.

1. **Conclusion**

Captivating conversations are taking place about the future of artificial intelligence and what it will/should mean for humanity. People`s opinion is very important for many businesses today. Technology is progressing really fast, and that scares some of them. Millennials are up for the change but there are still a lot of mix feeling particularly about Artificial Intelligence. There are fascinating controversies where the world’s leading experts disagree, such as: AI’s future impact on the job market; if/when human-level AI will be developed; whether this will lead to an intelligence explosion; and whether this is something we should welcome or fear. But there are also many examples of boring pseudo-controversies caused by people misunderstanding and talking past each other.

Experts say the rise of artificial intelligence will make most people better off over the next decade, but many have concerns about how advances in AI will affect what it means to be human, to be productive and to exercise free will.

**Sources:**

<http://www.nltk.org/howto/sentiment.html>

<https://levelup.gitconnected.com/sentiment-analysis-using-machine-learning-python-9122e03f8f7b>

<https://monkeylearn.com/sentiment-analysis/>

<https://www.aithority.com/guest-authors/strong-ai-doesnt-make-humans-less-special/>

<http://www.nltk.org/book/ch01.html>

<https://www.guru99.com/tokenize-words-sentences-nltk.html>

<http://t-redactyl.io/blog/2017/04/using-vader-to-handle-sentiment-analysis-with-social-media-text.html>

<https://towardsdatascience.com/fine-grained-sentiment-analysis-in-python-part-1-2697bb111ed4>

<https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/>