CHRIST
(DEEMED TO BE UNIVERSITY)
PUNE LAVASA CAMPUS
The Hub of Analytics

Kumar Gaurav 20212065

NLP assignment

Multilingual Public sentiment opinion polling

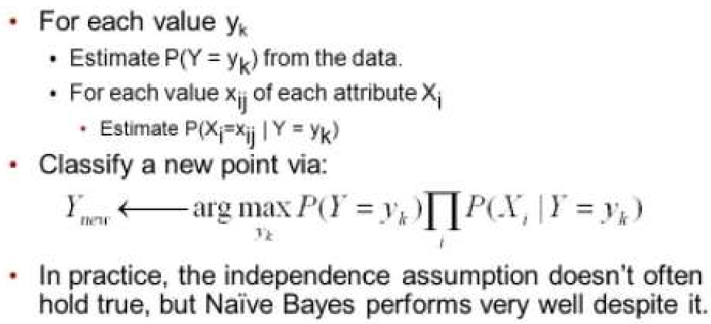
Kumar Gaurav, MSc (Data Science)

Abstract: An opinion poll, often simply referred to as a poll or a survey, is a human research survey of public opinion from a particular sample. Opinion polls are usually designed to represent the opinions of a population by conducting a series of questions.Language detection is a natural language processing task of identifying the language a given document is written in. It is often the first step in a document processing pipeline. Moreover, it is considered to be a critical pre-processing step in applications that require language specific modelling, such as search engines, where depending on the detected language different.A language usually refers to the spoken language, a method of communication. A script refers to a collection of characters used to write one or more languages. A language is a method of communication. Scripts are writing systems that allow the transcription of a language, via alphabet sets.Many language has same alphabet as English and the script has also same as English .That is challenge for us to detect this. You can measure online sentiment using sentiment analysis. Sentiment analysis often uses artificial intelligence to identify the emotional tone of an online mention such as social media posts. We can use public sentiment in a sentence- The advantages are so great that public sentiment is always advancing in the same direction. I expect that, long after my time, with a change of public sentiment it will be decided to abandon this lottery.

Introduction: Human behaviour is the potential and expressed capacity (mentally, physically, and socially) of human individuals or groups to respond to internal and external stimuli throughout their life-from Wikipedia. Human and human behaviour ply important role in Data Science. There are factors can affect behaviour- Physical factors - age, health, illness, pain, influence of a substance or medication. Personal and emotional factors - personality, beliefs, expectations, emotions, mental health. Life experiences - family, culture, friends, and life events. Human behaviour is learned in interaction with our environment, and that all behaviours are learned through experience. Behaviour Analysis is a realm of science that studies human learning and behaviour. According to this science, there are a set of rules, or principles that can be used to describe how these two things take place. Abnormal behaviour that is atypical or statistically uncommon within a particular culture or that is maladaptive or detrimental to an individual or to those around that individual. Opinion expressed by crowed on any social media sites. Learn more in: Analysis of Public Sentiments about Mega Online Sale Using Tweets on Big Billions Day Sale. The definition of a sentiment is a combination of beliefs and emotions that explains an action. An example of sentiment is someone being so patriotic that they decorate their house with many flags from their country. Appeal to the emotions in literature or art; expression of delicate, sensitive feeling.

Related works: If you don't consider "ij" as a separate letter, Dutch has the exact same alphabet as English. I think the African Language Xhosa also uses the same alphabet.Tokenizers may be used. Another common example of applying language detection is as a preceding step to machine translation, since the language of the text to be translated is not always specified. Therefore, a reliable language detection tool is needed. Many different approaches for tackling the language detection problem have been developed so far. Some of the best known models include the one of Cavnar and Trenkle [9] popularized in the textcat tool, the Chromium Compact Language Detector 2 (CLD2). CLD2 and langid are both Naive Bayes classifiers, where CLD2 probabilistically detects over 80 languages in Unicode UTF-8 text and for the mixed-language input returns the top three languages found for a given input and their approximate percentages of the total text bytes, while langid is 2trained on 97 languages over a naive Bayes classifier with a multinomial event model over a mixture of byte n-grams ($1 \leq n \leq 4$) designed to be used off-the-shelf [11]. In the Results Section, the performance of CLD2 and langid is compared to the performance of the methods developed in this paper

Challenges:I have to observe about overall human opinion and polling. If any celebrity, international org, prime minister or international face comments something or posted something or tweet something then I saw that they got lots



- For each value y_k
  - Estimate $P(Y = y_k)$ from the data.
  - For each value x_{ij} of each attribute X_i
    - Estimate $P(X_i=x_{ij} \mid Y = y_k)$
- Classify a new point via:

$$Y_{new} \longleftarrow \arg\max_{y_k} P(Y = y_k)\prod_i P(X_i \mid Y = y_k)$$

- In practice, the independence assumption doesn't often hold true, but Naïve Bayes performs very well despite it.

of reply and comments from different cultural, group and different country and differ language.

The problem is that, overall public opinion should be same. Sentiment polling from different language and different societies May be differ what I am thinking.

What will be final exact sentiment conclusion about this opinion that commented in many languages? So, this project tell about overall public opinion about how happy people are with them? Because people here come from different societies, different languages? The problem is that, I expect from someone, who care my post /view my pics but he didn't response that like as I wish.That time, my behaviour may be change toward his. A hatred developed in my heart for him. The problem is what I should do and is the real unbiased behave?We have to detect which language -Translation it into English, Tokenize it and working with NLP,Sentiment analysis of content,Overall opinion polling of content ,Visualize it

Methods: In this paper, two different types of features are extracted from the tweet texts, depending on the classifier used: character n-grams and bag-of-words features. Character n-grams can be described as all character substrings of length n in the given text. On the other hand, the bag-of-words features are defined as an unordered collection of words in the text.

$$w_{i,j} = tf_{i,j} \cdot log\left(\frac{N}{df_i}\right)$$

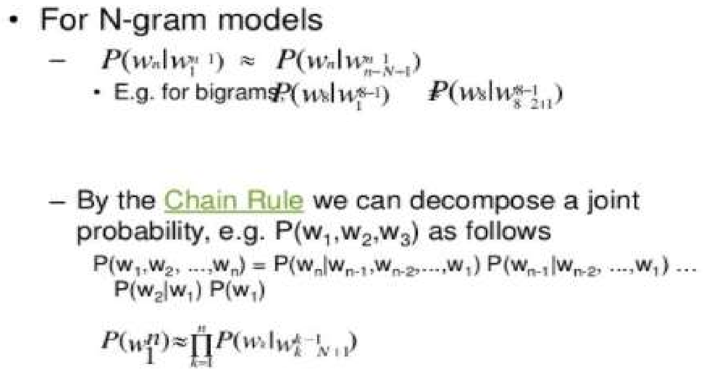
$tf_{i,j}$ = number of occurrences of i in j

$df_{i,j}$ = number of documents

N = total number of documents

Whenever possible, the character n-gram feature model is chosen over the bag-of-words model, which is justified by the specific type of language used in the dataset. Namely, character n-grams model is more resilient against misspellings, abbreviations, acronyms, and word derivations than the bag-of-words, since it does not strictly impose the splitting of texts by white spaces. For SVM and logistic regression classification, character n-grams are chosen as the appropriate feature type. After extracting the n-grams, the next step is to transform this collection of features into numerical feature vectors, which is a standard step before applying most of the machine learning algorithms to text data.

For detecting language purpose, I downloaded dataset from kaggle . I am working on training dataset and use Naïve Bayes algorithm, then I get 97 percent accuracy. For translation purpose , I use google translation python module ,Import google translation With Help with NLP , I will tokenize it ,bags of word , stem, lemma, And sentiment work ,Overall opinion – happy , sad, angry , curious etc.



- For N-gram models
  - $P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-N+1}^{n-1})$
    - E.g. for bigrams $P(w_8|w_1^{8-1})$ $P(w_8|w_{8-2+1}^{8-1})$
  - By the Chain Rule we can decompose a joint probability, e.g. $P(w_1,w_2,w_3)$ as follows

    $P(w_1,w_2, \ldots,w_n) = P(w_n|w_{n-1},w_{n-2},\ldots,w_1)\, P(w_{n-1}|w_{n-2}, \ldots,w_1) \ldots P(w_2|w_1)\, P(w_1)$

    $P(w_1^n) \approx \prod_{k=1}^{n} P(w_k|w_{k-N+1}^{k-1})$

Conclusions: In this paper, different algorithmic approaches to language detection for short texts in social media are investigated. The first approach includes the use of the well-known classifiers such as SVM and logistic regression and the combination of both. The second approach is based on a probabilistic model with modified Kneser-Ney smoothing, with the extension in terms of including additional information specific to a single user. The last approach is a simple dictionary based method. When comparing the classification performance of all the algorithms, the probabilistic model outperforms the other methods.

References:

[1] Ivana Balaˇzevi´c1 , Mikio Braun1 , Klaus-Robert M¨uller1 “Language Detection For Short Text Messages In Social Media” arXiv:1608.08515v1[2016]

[2] S Padmaja1 and Prof. S Sameen Fatima “Opinion Mining and Sentiment Analysis –An Assessment of Peoples’ Belief: A Survey” [2013]

[3] Siaw ling lo, Erik cambria,Raymond chiong,David cornforth " Multilingual sentiment analysis: from formal to informal and scarce resource languages" [2017]

[4] Rachel Macreadie, Research Officer "Public Opinion Polls" [3,July 2011] ISSN 1836-7941 (Print) 1836-795X (Online)