Project choice

To justify the choice of my project I need to describe my background. I have previously graduated a undergraduate and master's degree in Marketing. This has developed my passion for research and as part of my study scheme I had to prepare surveys, apply them and analyse the gathered data in statistic analysis software such as SPSS (pretty graphs and statistics). Later, during my industrial placement while working with clients I felt that the company can really benefit from a system that could analyse client feedback(language processing). When asked to evaluate and choose from tens of major project suggestions I was naturally drawn to the research ones. The Analysis of the Historical Newspaper data was my first choice as it seemed to have more freedom regarding what direction I want to go on. There were several options: improving the API, improving the text using image processing, correcting and transforming the articles by removing unnecessary characters and checking spelling errors done by the scanning process or using the actual articles to obtain information about historical events.

The National Library of Wales has made available online the historical newspaper archive for research purposes. The entire collections has over .... newspapers from a period of over 100 years.

The articles contain news and events that extends over a variety of domains such as: science, events, sport, politics, crime, advertisement. These are not classified as the news are today in sections depending on theme and there is no evident connection between the type of article and the page it's found on.

During the initial stage I considered different options and ideas about how the data can be used. I knew my goal was to make it interesting not only for me to work on but also for a future user that might want to discover specific facts about events that occurred a long time ago.

I fiddled with ideas such as:

**Advertisements in Wales history -** analysis of all advertisements, what type products/services are more advertised? How it changed in time? What product services are mentioned in news articles;

**Scientists in Wales** - discover all the mentions of scientists, and work out information such as: are they in the news for their science or  something else (crime, business, politics etc), what kind of science, were they living in Wales, what kind of language was used to describe  them (positive or negative), did this reporting change over time, what words were most frequently used to describe them?  
**Crime Stories -** Separating only the crime specific articles and then developing an algorithm that can split them into different types of crime: theft, murder, fraud, bodily harm, no crime. After applying the model on all the articles, time allowing, I will apply some statistic and text processing techniques to find different connections between place, sex, year and then showing them to the user on a website.

Sentiment Analysis and formal news data - This was an interesting idea that I have researched but after reading more of the articles I realised it is a very difficult task for a human to realise weather a news article is positive or negative (for example text about a specific gathering ) and this would make it even more difficult for a computer.

The project will have different stages that would be divided by the techniques used in each:

1. Manual labelling and research
2. Text processing and cleaning
3. Database choice and structure
4. Machine Learning
5. Applying the model from the machine learning
6. Front end presentation of finings.

The initial data has its limitations.

1. The API was available only from the Computer science department and I was provided with a username and password specifically made for me to be able to access it from my own device on any network.
2. The API can currently be applied only to a limited amount of news articles and it's meant to be updated with 650k more ...
3. Some articles are written in Welsh
4. Some articles are on different columns/pages which means that the extraction into the API was made into separate records.