Literature review

* Can social media be reliably analysed to provide useful statistical data?

Nadkarni, Prakash M; Ohno-Machado, Lucila; Chapman, Wendy W. (2011) Natural language processing: an introduction. *Journal of the American Medical Informatics Association.* [Online] **18**(5), pp.544 - 551 [Accessed 19 October 2014] Available at:

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3168328/?tool=pmcentrez&report=abstract>

This source is similar to the work of Wang *et al*. (2013) as both sources detail NLP and how it can be used although this source describes NLP itself in far greater detail as it assumes the reader does not have in depth knowledge of NLP and therefore gives an introduction to the topic, it does though give an insight into how NLP applications are designed something which is invaluable to my answering of the sub question. It does not quite go into as much detail on an example of its use and include testing of this such as in the second source which is very useful for my answering of the sub question, therefore while the first source better introduces the concept of NLP which while it is necessary for my answering of the sub question the second gives a clearer example of its use and will be much more useful in how I intend to answer my sub question. The source gives an adequate amount of background information on NLP, its references are of the latest research being done on the topic at that time and in abundance and therefore will be an important reference that will help answer my sub question and from that the main research question.

Kumar, S., Hu, X., & Liu, H. (2014). A behavior analytics approach to identifying tweets from crisis regions. *Cover Art: In Proceedings of the 25th ACM conference on Hypertext and social media - HT ’14* [Online]. Santiago, Chile 1-4 September. New York: ACM. [Accessed 6 November 2014] Available at: <http://dl.acm.org/citation.cfm?id=2631775.2631814>

This source is similar to the work of Wang *et al*. (2013) as both sources give detailed explanations of how NLP can be used while not giving much background information on NLP itself as well as both sources mentioning how the quality of input or training data can affect the performance of such systems. However the first source provides much greater emphasis on its own testing giving qualitative evidence to its observations , this provides enough of an insight into the usefulness of NLP thanks the study of tweets that was carried out, the analysis of the results leading to their creation of a method of seeing the location of tweets and their testing of this method which showed the method was indeed adequate for the task and therefore I believe this will be a useful reference in answering my sub question.

Wang, A., Hoang, C., & Kan, M.-Y. (2013). Perspectives on crowdsourcing annotations for natural language processing*. Language Resources & Evaluation* [Online] **47**(1), pp.9 - 31 [Accessed 7 November 2014] Available at: <http://wk6kg9sd8m.search.serialssolutions.com/?ctx\_ver=Z39.882004&ctx\_enc=info%3Aofi%2Fenc%3AUTF8&rfr\_id=info:sid/summon.serialssolutions.com&rft\_val\_fmt=info:ofi/fmt:kev:mtx:journal&rft.genre=article&rft.atitle=Perspectives+on+crowdsourcing+annotations+for+natural+language+processing&rft.jtitle=Language+Resources+and+Evaluation&rft.au=Wang%2C+Aobo&rft.au=Hoang%2C+Cong+Duy+Vu&rft.au=Kan%2C+MinYen&rft.date=20130301&rft.pub=Springer+Netherlands&rft.issn=1574020X&rft.eissn=15740218&rft.volume=47&rft.issue=1&rft.spage=9&rft.epage=31&rft\_id=info:doi/10.1007%2Fs10579-012-9176-1&rft.externalDBID=n%2Fa&rft.externalDocID=2013\_10579\_47\_1

\_9176&paramdict=en-US>

The source is similar to the work of Kumar *et al*. (2014) as both sources assume the reader already have an understanding if NLP, as this source illustrates how it is being used to improve machine learning techniques by the use of annotation programs while the second source deals with the use of NLP determining locations in social media which is much more appropriate for answering my sub question however the first source is still useful for showing how NLP can be successfully used to provide useful data, it includes a large amount of testing on this and the data that the machine learning algorithms use to learn from is integral to its performance therefore this reference will be very important in answering the main research question and my sub question.

O’Hare, N; Murdock, V; (2013) Modeling locations with social media. *Information Retrieval.* [Online] **16**(1), pp.30 - 62 [Accessed 28 November 2014] Available at: <http://search.proquest.com.ezproxy.wlv.ac.uk/docview/1283116799?pq-origsite=summon>

This source is similar to the work of Kumar *et al*. (2014) as both sources deal with gathering useful information from a social media network in this case locations of messages or photos, however the second source uses Twitter as its data source which is a much better source of qualitative data due to it being much more text based that flickr which is used in the first source. The source creates a method for the task which is tested and the results analysed in order to concluded that their method is an improvement on the techniques being employed at the time, also the source backs itself up by including a wealth of references from the latest research on the topics covered, therefore I believe this source will help answer my sub question by providing me with more information on the methods to acquire statistical data from social media.

Han, B; Cook, P; Baldwin, T; (2013) Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology (TIST)*, [Online] **4**(1), pp.1 - 27 [Accessed 30 November 2014] Available at: <http://dl.acm.org.ezproxy.wlv.ac.uk/citation.cfm?id=2414430>

This source is similar to the work of Kumar *et al*. (2014) as both sources look at how to gather specific information from Twitter for the task at hand, however their methodologies are different as the second source looks at finding the location of a tweet from the text and information on the owner by using behavioural patterns in order to create a model of a tweet while the first source looks to help NLP in this task by creating a method that identifies and normalises lexical variants by creating a classifier to detect and tweak them for a more favourable result. The method is tested on a body of twitter and SMS data which is analysed to show that their method is a better option than state of the art techniques of this type in use at the time also the source includes a large amount of recent references for the subjects touched upon therefore I believe this source will help me answer my sub question.

Keselman, A; Rosemblat, G; Kilicoglu, H; Fiszman, M; Jin, H; Shin, D; & Rindflesch, T. C; (2010). Adapting Semantic Natural Language Processing Technology to Address Information Overload in Influenza Epidemic Management. *Journal of the American Society for Information Science and Technology.* [Online] **61**(12), pp.2531 - 2543 [Accessed 28 November 2014] Available at: <http://onlinelibrary.wiley.com.ezproxy.wlv.ac.uk/doi/10.1002/asi.21414/abstract>

This source is similar to the work of Wang *et al*. (2013) as both sources deal with how NLP can be used to sift through a huge amount of data and pick out only the parts needed for the task at hand however the second source is very general as it covers NLP as a whole while the first source looks at this problem specifically. It creates a solution using semantic NLP, automatic summarisation and visualisation which is tested and the results show that the method performs adequately for the task at hand, the work and conclusions in this paper are backed up by the research carried out in this field that is chronicled in the reference list as they are of the latest research, therefore this source will prove useful when answering my sub question.

Asur, S; Huberman, B. A; (2010). Predicting the Future with Social Media. *IEEE Computer Society.* [Online] **1**, pp.492 - 499 [Accessed 30 November 2014] Available at: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5616710>

This source is similar to the work of Han, B; Cook, P; Baldwin, T; (2013) as both sources use Twitter as their data source for the task at hand however the second source has a different methodology in that it uses lexical normalisation to help NLP gather specific social media text while the first source creates a method to predict the profits of a film before its release by their creation of a linear regression model as well as the sentiment analysis of tweets for improving prediction capability. This source is useful as although it does not directly implement NLP techniques it does deal with social media and prediction which is a major part of the main question which it will help answer also the techniques and how they were employed especially the analysis of tweets give me ideas on how I will answer sub question and ultimately the main question.

KIM, K.-H., & KIM, J.-H. (2011). Domain Independent Vocabulary Generation and Its Use in Category-based Small Footprint Language Model. *Advances in Electrical and Computer Engineering.* [Online] **11**(1), pp.77 - 84 [Accessed 8 December 2014] Available at: <http://www.aece.ro/abstractplus.php?year=2011&number=1&article=13>

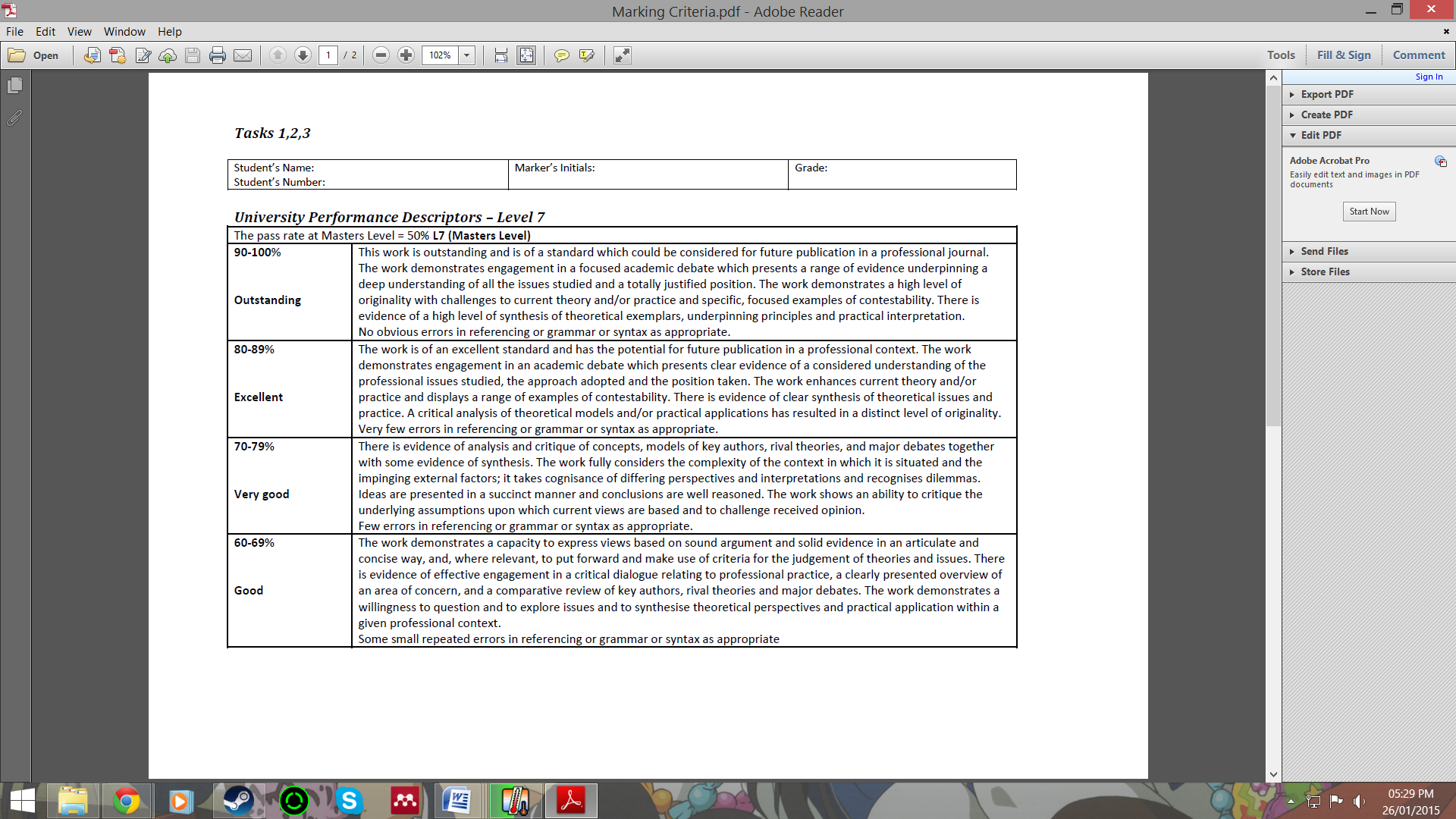
This source is similar to the work of Han, B; Cook, P; Baldwin, T; (2013) as although both sources use different methods they attempt to deal with the issue of NLP they also both test this using a body of SMS data. The second source uses a method to assist NLP using lexical normalisation with the creation of a classifier to tweak results favourably while the first source uses a vocabulary generation method and a small footprint language model for NLP and attempts to solve the issues of memory capacity and data sparsity, their method is tested against the body of SMS data and analysed to show it is an improvement in terms of coverage, memory capacity and performance over normal language models. Therefore this source is useful as it shows another side of NLP that I will take into account, it's references cover the topics touched upon well and back up its points and so this source will help me to answer my sub question.

Thom, D., Bosch, H., Krueger, R., & Ertl, T. (2014). Using Large Scale Aggregated Knowledge for Social Media Location Discovery. *2014 47th Hawaii International Conference on System Sciences*. [Online] **47**, pp.1464 - 1473 [Accessed 30 November 2014] Available at: <http://ieeexplore.ieee.org.ezproxy.wlv.ac.uk/xpls/abs\_all.jsp?arnumber=6758786&tag=1>

This source is similar to the work of Kumar *et al*. (2014) as both sources look at locations in social media however the second source uses a method that detects location in tweets merely from the text and any information on the owner the first source uses large scale aggregate knowledge in order to determine the location of twitter messages this provides me with even deeper knowledge of NLP and the options I have when it comes to answering my sub question and main question, also the source includes references from the latest research which are based on its problem and solution which effectively backs itself up therefore I believe this source will play an integral part in my answering of the sub and main question.

Zhou, D., Lawless, S., & Wade, V. (2012). Improving search via personalized query expansion using social media. *Information Retrieval*, [Online] **15**(3), pp.218 - 242 [Accessed 30 November 2014] Available at: <http://search.proquest.com.ezproxy.wlv.ac.uk/docview/1015641097?pq-origsite=summon>

This source is similar to the work of Asur, S; Huberman, B. A; (2010) as both sources deal with the use of social media and do not directly employ NLP techniques as the second source predicts how good of a profit a film makes using a linear regression model and sentiment analysis which is more appropriate for answering my sub and main question while the first source creates a query expansion framework based on user profiles in order to improve web search which is less appropriate but still deals with providing useful data from social media which is an integral part of answering my sub question.



James Braznell

1007022

