*E*ntity-level *E*vent Impact *A*nalytics

Govind Govind

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité INFORMATIQUE

Préparée au sein de l'Université de Caen Normandie

Entity-level Event Impact Analytics

Présentée et soutenue par
*** GOVIND**

Thèse soutenue publiquement le 12/12/2018 devant le jury composé de		
M. PATRICE BELLOT	Professeur des universités, Aix-Marseille Université	Rapporteur du jury
Mme BRIGITTE GRAU	Professeur des universités, ENSIIE	Rapporteur du jury
Mme CELINE ALEC	Maître de conférences, UNIVERSITE CAEN NORMANDIE	Membre du jury
M. PIERRE SENELLART	Professeur des universités, ECOLE NORMALE SUPERIEURE PARIS	Membre du jury
M. MARC SPANIOL	Professeur des universités, UNIVERSITE CAEN NORMANDIE	Directeur de thèse

Thèse dirigée par MARC SPANIOL, Groupe de recherche en informatique, image, automatique et instrumentation (Caen)

Présentation en français

Introduction

Les interactions entre des entités du monde réel telles que des pays, des groupes politiques, des organisations commerciales, etc. définissent le passé et façonnent l'avenir du monde. De ces interactions sémantiques naissent les événements. En effet, quand des entités sortent de leur routine habituelle, cela correspond en général à un événement, et est porteur d'une signification particulière. Disposer d'outils et de méthodes appropriés pour comprendre le lien entre des événements et des entités peut révéler des informations importantes sur l'état général de la société.

En intégrant les récents développements des technologies de l'information et de la communication, une grande partie de la population mondiale peut accéder à Internet, et par conséquent, de plus en plus de personnes génèrent des contenus sur le Web et/ou utilisent des contenus du Web. Aujourd'hui plus que jamais, le Web enregistre un nombre considérable d'informations d'actualité. Différentes plates-formes en ligne, telles que Twitter[1], font office à la fois de médias sociaux et de médias d'informations et enregistrent quotidiennement un volume considérable d'informations. L'Internet Archive[2] contient des archives d'environ 350 milliards de pages Web archivées à différents moments du temps depuis 1996. En dehors de cela, il existe d'innombrables agences de presse mondiales et locales, qui font le point sur les actualités de la société et rendent ces dernières disponibles en ligne. Chaque jour, plus de 500 nouveaux articles sont ajoutés à la version anglaise de Wikipedia, et plus de 100 modifications par minute[3] sont effectuées sur les articles existants (environ 6 millions). En outre, il existe plus de 200 versions de Wikipedia couvrant les principales langues du monde. La Figure 1.1 montre le lien entre les événements sociétaux du monde réel et ceux sur le Web. Le Web reflète chaque événement sociétal du monde réel en mentionnant les entités participantes et influencées par celui-ci. En conséquence, il serait juste de dire que le Web est le reflet du monde réel du point de vue des événements sociétaux. La numérisation des informations historiques et sociales à grande échelle et des activités quotidiennes ouvre la porte à beaucoup de possibilités de recherche. Mais cela pose également de nombreux problèmes, tels que la catégorisation automatique et l'organisation des contenus Web pour un meilleur stockage et une récupération ultérieure. Le Web enregistre des informations sur la société qui peuvent être utilisées pour de nombreuses tâches d'analyse sociale telles que la prévision d'un trouble

[1] https://www.twitter.com
[2] https://www.archive.org
[3] https://en.wikipedia.org/wiki/Wikipedia:Statistics

civil, la propagation d'une maladie épidémique, l'état des marchés financiers, etc. Ces informations peuvent également être utilisées pour analyser les événements de manière rétrospective de même que pour des tâches de prédiction, toutes aussi importantes voire plus importantes, qui s'intéressent à la question « Que vient-il ensuite? ».

Dans cette thèse, nous abordons des tâches de prédiction et des défis tels que la diffusion d'événements sociétaux dans les communautés de langue étrangère, la catégorisation automatisée de contenus Web et l'analyse de la viralité des informations en ligne. La tâche de prédiction de la diffusion d'événements découvre les communautés de langues étrangères dans lesquelles un événement se propagera à l'avenir, en fonction des données de l'événement collectées au cours d'une courte période initiale. Nous abordons également le problème de la classification automatisée des contenus Web, qui consiste à aligner chaque contenu Web avec une hiérarchie de types assez précis. Enfin, nous explorons la prédiction automatisée de la viralité et de la pertinence des articles de presse en ligne par rapport aux régions géographiques. Dans les sections suivantes, nous discutons de l'analyse des événements de société et nous fournissons l'intuition derrière le fait d'exploiter les connaissances au niveau des entités afin de mieux comprendre un document.

L'analyse des événements sociétaux

L'analyse des événements sociétaux s'intéresse à la détection et à l'analyse d'événements, ce qui revêt une grande importance pour la société. En particulier, nous considérons un événement comme étant « pertinent pour la société » si la communauté d'utilisateurs de Wikipedia crée une entrée correspondante. Un événement peut être décrit par un ensemble d'entités ayant une activité inhabituelle par rapport à leur comportement normal. Comme les entités du monde réel peuvent avoir une influence et une réputation variables, la dynamique et l'évolution d'un événement les impliquant sont également affectées.

L'analyse de l'impact des événements

Les événements sociétaux ont un impact et des conséquences sur l'avenir. Dans un tel scénario, connaître la manière dont un événement va se propager dans l'espace temporel et géographique revêt une grande valeur. L'impact d'un événement se reflète sous diverses formes en fonction des caractéristiques de celui-ci, par exemple, une panne de serveur informatique due à un trafic important provoqué par un événement sociétal, ou à des pertes catastrophiques causées par des catastrophes naturelles, etc. La connaissance du déroulement/de la propagation d'un événement permet aux autorités respectives de se préparer de manière proactive à une meilleure gestion de l'impact. La propagation d'un événement peut être facilement imaginée dans les dimensions spatiale et temporelle, mais étant donnée l'ère actuelle de l'information, un événement peut se propager par d'autres canaux de diffusion, tels que parmi les communautés de langues sur le Web.

De nos jours, les événements sociétaux ont une diffusion presque mondiale en raison de la croissance considérable de la couverture des articles de presse en provenance de pratiquement toutes les régions du monde. Par exemple, les utilisateurs mondiaux de

Facebook n'ont qu'un degré de séparation de 3,5 [4], ce qui implique que si un événement présente un intérêt significatif pour une communauté d'utilisateurs particulière, il parviendra facilement à atteindre cette communauté. L'ampleur de la diffusion varie beaucoup en fonction de l'ampleur de l'incident. Par exemple, un événement de taille relativement réduite `Oscars 2017 Mix-up` dans un contexte hyper-local a moins de chances d'avoir une diffusion globale que celui de `Executive Order 13769` alias « Muslim Travel Ban »[5] qui concerne un public beaucoup plus large. Un incident mineur pourrait se limiter aux réactions à court terme sur les réseaux sociaux uniquement, tandis qu'un événement concernant un public plus large aura un impact considérable. Ainsi, un événement sociétal pertinent pourra être relaté dans une plate-forme encyclopédique collaborative, telle que Wikipedia. On peut s'attendre à ce que l'impact d'un événement sociétal pertinent ne soit pas nécessairement lié à une seule communauté. Par exemple, le `Executive Order 13769` susmentionné a été traduit dans environ 25 langues en plus de la langue d'origine, c'est-à-dire l'anglais. Celles-ci sont composées de plusieurs des langues principalement parlées dans les pays affectés, notamment l'arabe et le persan, qui sont également parmi les premières à avoir obtenu des traductions.

Cependant, la question clé est la suivante : comment évaluer automatiquement l'impact des événements sociétaux pertinents en ce qui concerne leur potentiel de diffusion dans les communautés de langue étrangère? Nous émettons l'hypothèse que les entités impliquées dans un événement possèdent une connaissance sémantique qui fournit de puissants indices contextuels pour prédire avec succès la diffusion d'événements dans des communautés de langues étrangères. Dans la sous-section suivante, nous discutons de l'importance des entités dans l'analyse de l'impact des événements et dans la compréhension approfondie des contenus Web en général.

L'importance contextuelle des entités

L'évolution et la trajectoire des événements sont sensibles à leur interaction avec les entités du monde réel. Dans plusieurs exemples d'événements sociétaux tels que `Hurricane Katrina`, `WannaCry Ransomware`, `Executive Order 13769`, etc., nous remarquons la même sensibilité. Par exemple, dans le cas du `Executive Order 13769`, les régions géographiques concernées par cet événement sont fortement influencées par les entités qu'il contient. Le sac-de-mots est l'une des représentations de documents les plus couramment utilisées pour plusieurs tâches de prédiction. Il consiste à représenter un document par les mots qu'il contient et à calculer leur fréquence. Il n'utilise pas d'entités nommées canonique et ignore toute information au niveau des entités. Ainsi, il sera difficile de saisir la sémantique du document en utilisant une représentation en sac-de-mots car celle-ci ne permet de capturer une connaissance sémantique au niveau des entités. Cela soulève la question suivante : comment représenter ou traiter un document pour en saisir la sémantique et mieux le comprendre? Pour ce faire, une interprétation humaine est nécessaire pour faciliter la réalisation de divers objectifs de niveau supérieur, tels que la prévision de la diffusion d'événements au sein de certaines communautés ou la classification de docu-

[4] https://research.fb.com/three-and-a-half-degrees-of-separation/
[5] https://en.wikipedia.org/wiki/Executive_Order_13769

ments dans des catégories détaillées. Alors que les techniques de traitement de la langue naturelle progressent bien, la compréhension/l'interprétation sophistiquée d'un contenu en est encore à ses balbutiements. Cependant, des indices importants peuvent encore être dérivés de la « sémantique inhérente » d'un contenu. Notre approche vise donc à exploiter des techniques d'analyse au niveau des entités afin d'effectuer diverses tâches de prédiction.

Élever un contenu au niveau des entités présente plusieurs avantages. L'un de ces avantages est la possibilité d'incorporer diverses connaissances issues de bases de connaissances. De nos jours, il existe des sources de connaissances à grande échelle librement accessibles contenant une mine d'informations sur les entités nommées, telles que DBpedia [Auer et al., 2007] ou YAGO [Suchanek et al., 2007, Hoffart et al., 2013]. De plus, avec l'émergence d'outils permettant d'interconnecter des documents textuels avec le Web des données (LOD), tels que AIDA [Hoffart et al., 2011, Yosef et al., 2011] ou DBpedia Spotlight [Mendes et al., 2011], il existe des moyens efficaces d'augmenter la sémantique d'un texte brut en l'élevant au niveau des entités. Ce faisant, nous sommes en mesure d'exploiter les connaissances structurées stockées dans des bases de connaissances et de tirer de précieuses informations sur les entités impliquées, ce qui peut aider à obtenir de meilleures performances pour les tâches de prédiction de haut niveau susmentionnées.

Approche et contributions

Les principales contributions de cette thèse sont au nombre de trois. La première est la prédiction de la diffusion des événements. Pour ce faire, nous présentons un framework qui exploite la sémantique au niveau des entités et prédit la diffusion des événements dans les communautés de langues étrangères. La seconde est la classification des contenus, pour laquelle nous proposons une nouvelle approche de représentation sémantique des documents exploitant les entités, nommée « l'empreinte sémantique ». La dernière contribution est l'évaluation et la visualisation de la viralité des articles de presse. Une implémentation en ligne de ce travail est déployée sur le Web.

Chacune des trois tâches susmentionnées est intrinsèquement difficile en raison de sa nature complexe. Il n'est pas facile de prédire la diffusion d'un événement car cela nécessite une compréhension plus approfondie des diverses caractéristiques de l'événement. En outre, l'alignement des contenus Web avec une hiérarchie détaillée exige de saisir les délicates différences de nature entre différents types. Pour prédire la viralité des articles de presse, il faut découvrir le lien sémantique et capturer la perception sociale. Afin de remédier à cela, l'approche conceptuelle de notre travail est centrée sur les entités. Nous élevons le contenu d'un document au niveau des entités afin de mieux saisir sa sémantique. Cela permet d'acquérir une compréhension sémantique plus profonde qui peut être utilisée pour une variété de tâches de prédiction. La Figure 1.2 donne un aperçu de l'approche conceptuelle de notre travail. On peut voir que les étapes initiales impliquent l'extraction et la désambiguïsation des entités nommées, suivies de l'analyse au niveau des entités utilisant les connaissances sociétales distillées et dérivées des bases de connaissances. Par la suite, les documents sémantiquement augmentés sont utilisés par des algorithmes de prédiction spécifiques à chaque tâche afin d'atteindre des objectifs disparates. Nous fournissons un bref aperçu de chacune des contributions susmentionnées

dans les sous-sections suivantes.

La prédiction de la diffusion d'un événement

La tâche de prédiction de la diffusion d'un événement s'intéresse à la prédiction et à l'analyse de la diffusion des événements sociétaux via les informations collectées dans le cyber-espace. Cela revêt une grande importance. En effet, cela fournit des indications sur l'évolution future des événements quotidiens de la société. Les informations préliminaires sur l'évolution d'un événement peuvent être utilisées à plusieurs fins, telles que l'allocation appropriée des ressources pour s'adapter à la propagation des événements, etc. Afin d'enquêter sur la diffusion d'événements imprévus (catastrophes naturelles, conflits, actions politiques, etc.), nous présentons le framework ELEVATE (Entity-LEVel AnalyTics for Event diffusion prediction). Le framework ELEVATE effectue une analyse sémantique des contenus Web en utilisant les entités contenues et les ressources du LOD (ici, la base de connaissances YAGO [Suchanek et al., 2007]). Contrairement aux approches existantes, notre approche exploite les connaissances des entités mentionnées, via des bases de connaissances. Après l'agrégation systématique de ces connaissances, un modèle de prédiction de la diffusion des événements est formé, qui apprend à prédire la diffusion d'un événement dans des communautés de langues étrangères. En résumé, les principales contributions de ce travail sont les suivantes:

- la définition d'un modèle de prédiction de l'impact d'un événement,
- l'utilisation du Web des données (LOD) pour analyser (sémantiquement) des contenus Web,
- le test avec un classifieur multi-étiquettes pour identifier des motifs de diffusion afin d'améliorer le rappel tout en maintenant la précision,
- une étude expérimentale complète sur les événements de Wikipedia couvrant une période de presque deux décennies et démontrant la qualité de notre méthode.

Classification des contenus au niveau des entités

La classification de contenus consiste à attribuer des classes/types/catégories approprié(e)s aux contenus Web ou aux documents en général. Elle s'applique à de nombreux cas d'application, qui varient en fonction du nombre de classes à classifier. Dans ce travail, nous nous concentrons sur le problème de l'alignement d'un contenu avec une hiérarchie de types très précis. Conformément à l'idée fondamentale de cette thèse consistant à exploiter la sémantique via des entités, nous introduisons « l'empreinte sémantique » comme une nouvelle approche de classification précise de contenu exploitant les entités. L'approche d'empreinte sémantique capture la sémantique inhérente d'un contenu Web en s'intéressant au(x) type(s) des entités contenues, à l'aide de la base de connaissances YAGO. Dans ce but, nous étudions la classification des contenus Web dans une hiérarchie de types avec une granularité fine (> 100 types) en exploitant la sémantique des entités. Une partie essentielle de ce travail consiste à étudier la classification des événements en une vingtaine de types précis. Les principales contributions de ce travail sont les suivantes:

- un nouveau modèle de représentation des contenus Web nommé « Empreinte Sémantique »,
- une approche basée sur de l'apprentissage automatique qui permet une classification fine des contenus Web grâce à des « empreintes sémantiques »,
- une étude complète sur la classification de contenus Web basée sur le système de catégorisation de Wikipedia, qui montre l'avantage de notre approche par rapport à des concurrents de l'état de l'art.

Évaluation et visualisation de la viralité des articles de presse

La prédiction de la viralité des articles de presse consiste à évaluer automatiquement la future popularité d'un article de presse. Dans ce travail, nous présentons le système ELEVATE-live, qui est une extension de notre framework de prédiction de diffusion d'événements ELEVATE [Govind and Spaniol, 2017]. ELEVATE-live est un système de prédiction de la viralité des articles de presse qui permet aux utilisateurs d'explorer et de visualiser la viralité/pertinence des articles de presse en ligne en ce qui concerne les lieux géographiques (en particulier les pays). En résumé, notre travail apporte les contributions suivantes:

- intégration du framework ELEVATE et élévation des contenus Web au niveau des entités pour l'analyse sémantique,
- exploitation de bases de connaissances afin de révéler des interdépendances non triviales entre les entités nommées contenues et les pays associés,
- fourniture d'une interface Web pour étudier la « viralité » des articles de presse par rapport aux pays concernés et vice versa.

Structure de la thèse

La structure générale de la thèse est organisée comme suit. Dans le Chapitre 2, nous fournissons une introduction détaillée aux principes fondamentaux et au contexte technique nécessaires pour comprendre le travail réalisé dans le cadre de cette thèse. Nous expliquons le principe du LOD et des bases de connaissances. De plus, nous présentons les concepts fondamentaux des modèles d'apprentissage automatique et discutons en détail de ceux qui sont utilisés dans notre travail. Nous abordons également certaines des méthodes d'évaluation largement utilisées, qui conviennent bien aux tâches examinées ici.

Ensuite, dans le Chapitre 3, nous discutons en détail des travaux proches de notre problématique. Nous résumons et comparons diverses études d'analyse d'événements sociaux allant de la détection d'événements à l'analyse d'impact, en passant par la prédiction et l'analyse de la viralité. Nous examinons ensuite des travaux disparates basés sur des modèles de base de connaissances et des analyses exploitant les entités. Nous concluons ce chapitre en discutant des études récentes sur la tâche de classification par type. Nous

abordons à la fois la classification par type d'entité et la classification par type de document, mais nous nous concentrons davantage sur les travaux au niveau du document car ceux-ci sont plus pertinents pour notre travail.

Dans le Chapitre 4, nous présentons notre travail sur la diffusion d'événements dans les communautés de langue étrangère. Plus précisément, nous présentons le framework ELEVATE (Entity-LEVel AnalyTics for Event diffusion prediction). Nous rapportons les résultats de nos expériences approfondies menées sur la tâche de prédiction de la diffusion des événements sur une centaine de langues. Nous prenons en compte plusieurs bases de référence de l'état de l'art et de multiples variantes de notre modèle. De plus, nous étudions l'évolution de diverses mesures de performance en effectuant des expériences à trois temps différents après le début d'un événement. Nous introduisons un modèle d'apprentissage automatique qui apprend les motifs de diffusion typiques et améliore les performances. Enfin, nous résumons nos résultats globaux issus d'expériences sur la tâche de prédiction de la diffusion des événements.

Ensuite, au Chapitre 5, nous abordons la tâche de classification des contenus utilisant des types précis. La découverte des types précis pour un événement et pour un contenu Web en général est une tâche cruciale en organisation des informations, car elle sert de germe à de nombreuses tâches de niveau supérieur. Nous proposons « l'empreinte sémantique », une approche qui fournit une représentation sémantique concise et efficace des documents basés sur les entités contenues. Nous utilisons l'approche proposée pour aligner les contenus Web avec une hiérarchie fine de 105 types, composée de 20 sous-types les plus fréquemment utilisés pour chacun des 5 types de niveau supérieur (événement, organisation, lieu, artefact, personne). Nous rapportons nos résultats sur un jeu de données de Wikipédia et nous comparons avec plusieurs bases de référence de l'état de l'art.

Dans le Chapitre 6, nous examinons la tâche d'analyse de la viralité des informations sous son angle géographique. Nous présentons ELEVATE-live, une extension du framework ELEVATE permettant d'évaluer la viralité des articles de presse en ligne. ELEVATE-live fournit un mécanisme permettant d'évaluer et de visualiser la pertinence d'un article d'actualité par rapport à des emplacements géographiques. Dans ce but, une interface en ligne a été créée et déployée dans le cadre d'ELEVATE-live. Celle-ci est disponible gratuitement sur le Web pour un usage public. Dans le cadre de ce travail, nous surveillons diverses agences de presse en ligne réputées et permettons à l'utilisateur final de vérifier la viralité des informations en temps réel. Nous fournissons la visualisation à l'aide d'une carte du monde géographique et d'une chronologie. De même que le thème général de la thèse, ELEVATE-live est également basé uniquement sur une analyse exploitant des entités.

Enfin, dans le Chapitre 7, nous concluons l'ensemble du travail accompli et exprimons des possibilités de recherche futures. Nous résumons chaque résultat chapitre par chapitre, mais dans ce dernier chapitre, nous examinons le travail de manière plus globale et proposons une perspective plus large.

Publications dans le cadre de cette thèse

Les travaux de cette thèse ont été publiés lors de conférences internationales renommées. Le framework ELEVATE a été publié dans la 9^{th} ACM Web Science Conference 2017 [Govind and Spaniol, 2017]. Nous avons reçu une subvention de voyage pour présenter le framework ELEVATE. La Web Science Conference est l'un des lieux privilégiés pour aborder les méthodes et les résultats permettant de développer notre compréhension du Web. La conférence a pour objectif de réunir les chercheurs travaillant dans plusieurs disciplines telles que l'informatique et les sciences de l'information, la sociologie, la psychologie, etc. L'importance des entités dans la compréhension et la représentation de documents basée sur « l'empreinte sémantique » pour la classification de contenus est publiée dans l'International Conference on Web Engineering (ICWE) 2018 [Govind et al., 2018b]. Le framework ELEVATE-live est également publié dans ICWE 2018 [Govind et al., 2018a]. La conférence ICWE est un lieu réputé pour la recherche sur les défis émergents en matière d'ingénierie d'applications Web et sur l'impact social et culturel du Web. De plus, pour présenter un poster [Govind, 2018] sur ce travail de thèse, une bourse a été attribuée par la WSNET Web Science Summer School 2018.

Govind and Spaniol, M. (2017). ELEVATE: A Framework for Entity-level Event Diffusion Prediction into Foreign Language Communities. In Proceedings of the 9th International ACM Web Science Conference (WebSci'17), pages 111–120.

Govind, Alec, C., and Spaniol, M. (2018b). Semantic Fingerprinting: A Novel Method for Entity-Level Content Classification. In Proceedings of the 18th International Conference on Web Engineering, ICWE 2018, Caceres, Spain, June 5-8, 2018, pages 279–287.

Govind, Alec, C., and Spaniol, M. (2018a). ELEVATE-Live: Assessment and Visualization of Online News Virality via Entity-Level Analytics. In Proceedings of the 18th International Conference on Web Engineering, ICWE 2018, Caceres, Spain, June 5-8, 2018, pages 482–486.

Govind (2018). Entity-level Event Impact Analytics. WSTNET Web Science Summer School, WWSSS 2018, Hannover, Germany, July 30 - Aug 4, 2018.

Conclusion

Comme mentionné précédemment, notre société observe une présence virtuelle croissante sur le Web, ce qui permet d'utiliser diverses méthodes de calcul pour étudier les événements de société. Cette thèse se focalise sur la compréhension d'événements sociétaux décrits sur le Web et examine leurs différents aspects tels que la diffusion, le type, la viralité, etc. Un aspect remarquable des méthodes proposées ici est qu'elles exploitent de manière extensive les informations sémantiques dérivées des entités mentionnées dans les descriptions des événements. Nos expériences permettent de constater que l'utilisation d'analyses au niveau des entités permet de mieux comprendre les documents du Web et qu'un ensemble diversifié de tâches peut utiliser ces connaissances pour s'améliorer.

Abstract

Our society has been rapidly growing its presence on the Web, as a consequence we are digitizing a large collection of our daily happenings. In this scenario, the Web receives virtual occurrences of various events corresponding to their real world occurrences from all around the world. Scale of these events can vary from locally relevant ones up to those that receive global attention. News and social media of current times provide all essential means to reach almost a global diffusion. This means that if a societal event is of significance to someone, then it is likely to find its way to them. This big data of complex societal happenings provide a platform to the multitude of research opportunities for analyzing and gaining insights into the state of our society. In this thesis, we investigate a variety of social event impact analytics tasks. Specifically, we address three facets in the context of events and the Web, namely, diffusion of events in foreign languages communities, automated classification of Web contents, and news virality assessment and visualization. We hypothesize that the named entities associated with an event or a Web content carry valuable semantic information, which can be exploited to build accurate prediction models. To this end, we study several research tasks as discussed in the following.

A broad level impact of an event can be assessed as being proportional to the volume of attention it receives on the Web. However, this does not provide any knowledge about communities that are subjected to the impact from an inter-cultural perspective. Inspired by its relevance, we investigate the task of predicting societal event diffusion into foreign language communities. Driven by our broader hypothesis, we study the societal event diffusion via entity-level knowledge. To accomplish that, we introduce the ELEVATE framework, which performs entity-level analytics on Web contents and incorporates semantic knowledge via Linked Open Data. We perform a thorough study on events ranging around two decades over temporal dimension. Our approach achieves very encouraging results and have effectively shown its advantages over the competitor methods.

Subsequently, we investigate that, whether entity-level analytics can be utilized to characterize a Web content. To this end, we explore the problem of aligning Web contents to a fine-grained type hierarchy. For this purpose, we introduce the "semantic fingerprinting" method that distills a Web content to retrieve its semantic representation based on the associated named entities. It encodes the specificity as well as the generality about the semantic nature of a Web content into a concise vector. Further, this semantic representation vector can be utilized for various tasks. Our accomplished experiments on type prediction task have shown that semantic fingerprinting achieves a deeper understanding of Web contents, and outperforms the state-of-the-art competitors when used in combination with machine learning.

Finally, we explore the task of news virality analytics. This study is driven by the need of automatic assessment for news virality, and to provide an effective mechanism for its visualization. We address the geographical aspect regarding the virality/relevancy of online news articles. To this end, we introduce ELEVATE-live, which is an extension of ELEVATE. Our approach exploits the entity-level information in order to harness semantic connections between news articles and different geographical regions. In order to accomplish this, we developed a Web-based news virality analytics and visualization platform that is available online for general public. The geo-temporal enabled interface of ELEVATE-live has successfully proven its efficacy.

In this thesis, we have shown with the help of multiple studies that raising Web contents to the entity-level captures their core essence, and thus, provides a variety of benefits in achieving better performance in diverse tasks. We report novel findings over disparate tasks in an attempt to fulfill our overall goal on societal event impact analytics.

Acknowledgment

First of all, I would like to express my sincere gratitude to my advisor, Prof. Dr. Marc Spaniol, for his mentorship and encouragement that made this dissertation possible. I would like to thank him for his guidance, patience, and welcoming gestures for new ideas. I appreciate his positive criticism that has thrived me to do better. I will always cherish the learnings from him.

I would like thank Dr. Céline Alec for her noteworthy contribution to my professional time at the GREYC. I am thankful for the simulating discussions, her precious feedback, and help with the French language.

Special thanks goes to Prof. Brigitte Grau, Prof. Pierre Senellart, and Prof. Patrice Bellot for accepting to be the part of my thesis defense jury. I would like to acknowledge their valuable feedback and positive criticism that served as a great help.

I am thankful to GREYC - UMR 6072 laboratory and Université de Caen Normandie for financially supporting my research. I would like to acknowledge the system administrators at the laboratory whose quick technical support kept the experiments running smoothly. I would like to thank Arielle Perrette and others for their patience and being helpful through the administrative works.

I would like to thank Emmanuel Giguet, Loïs Vanhée and all the members of team HULTECH for their friendly gestures that have always been uplifting and motivating.

Last but not the least, I would like to thank my parents and sisters for always being a source of encouragement. I would like to thank my wife, Anupriya, for being so caring and supportive through good and bad. I would also like to thank my friends, Amit, Dinesh, Ravinder and Deepak for their constant support in this journey.

Contents

Chapter 1

Introduction

1.1 Motivation and Problem

Interaction among real world entities such as countries, political groups, business organization, etc. define the past and shape the future of the world. From these semantic interaction originate events. Events represent the important activities from interaction among entities, which carry higher significance than the routine chatter. Having suitable tools and methods to understand this interplay between events and entities can reveal important insights about the overall state of the society.

Incorporating the recent developments in information and communication technology (ICT), the Internet has gained wide access to the world population and thus, more people are generating and consuming Web contents. Now more than ever, the Web records a tremendous number of activities from everyday happenings in the society. Various online platforms such as Twitter[6], have the characteristics of both social and news media, and record a huge volume of information on a daily basis. The Internet Archive[7] has

[6] https://www.twitter.com
[7] https://www.archive.org

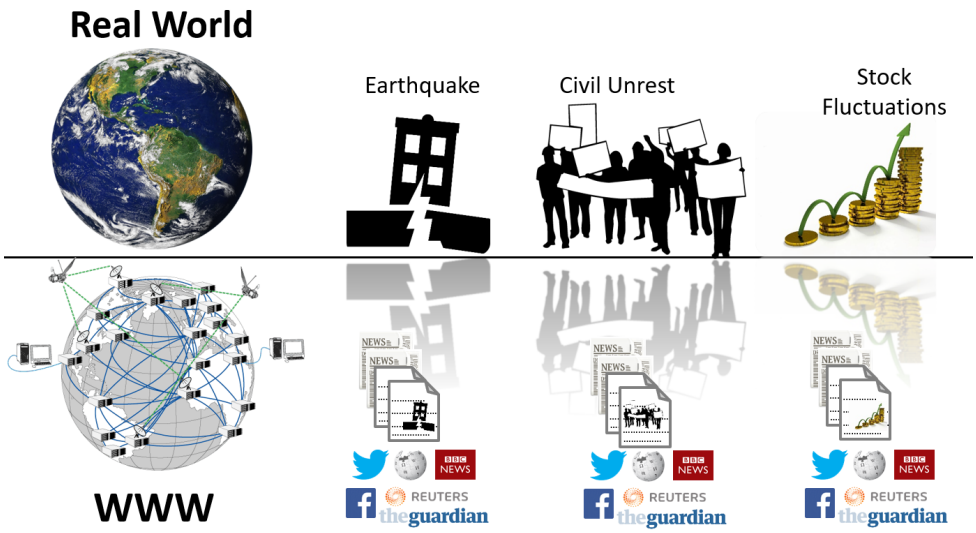


Figure 1.1: Reflection of societal events from real world on the Web

records of about 350 billion Web pages that have been archived at different time points along the temporal dimension since 1996. Apart from that, there are countless global and local news agencies, which report about the on-going activities in our society and make them available online. Interestingly, more than 500 new articles are added to the English version of Wikipedia everyday, and over 100 edits per minute[8] are performed on the existing articles (currently around 6 million articles). Moreover, there are more than 200 versions of Wikipedia covering major languages of the world. Figure 1.1 illustrates the notion of interplay between societal events in real world and on the Web. For every societal happening in the real world, the Web receives reactions in terms of data generated by the participating and influenced entities. As a result, it would be fair to say that the Web is reflection of the real world from the perspective of societal events. It has also been observed that online activities have the potential of affecting "on the ground" developments in events [Bastos et al., 2015]. The digitization of this large-scale historical, social information, and daily activities, open doors to a plethora of research opportunities. However, it raises many challenges too, such as automatic categorization and organization of Web contents for better storage and later retrieval. The Web records information about different aspects of society that can be utilized for many social analytics tasks such as the prediction of civil unrests, spread of epidemic diseases, state of financial markets, etc. This information can be used to analyze events retrospectively as well as for the equally or more important prediction tasks that address "what is coming next?".

In this thesis, we address various prediction tasks and challenges such as the diffusion of societal events into foreign language communities, automated Web content categorization, and virality analysis of online news. The event diffusion prediction task discovers foreign language communities into which an event will flow in the future, based on data harvested within a small initial period of event's origin time. Further, we take our research to address the problem of automated Web content classification, where we align Web contents with

[8] https://en.wikipedia.org/wiki/Wikipedia:Statistics

a fine-grained type hierarchy. Finally, we explore the automated prediction of virality and relevance of online new articles towards geographical regions. In the following sections, we discuss societal events analytics and provide an intuition behind harnessing knowledge at the entity-level in order to gain a deeper semantic understanding.

1.2 Societal Event Analytics

Societal event analytics deal with the detection and analysis of events, which are of high significance to the society. Particularly, we consider an event as "societal relevant" if a corresponding entry is created by the user community of Wikipedia. An event can be broadly modeled by a collection of co-occurring entities, which are having unusually high activity when compared to their normal behavior. As real world entities have varying influence and reputation, respectively the dynamics and the evolution of an event associated with them, are also affected. A societal event (for example `Hurricane Katrina`[9]) evolves over time and space, and can develop various characteristics [Dos Santos et al., 2016] such as the following.

- **Event cascading:** On its own, an isolated event provides only a limited insight into the overall picture. For example, `public evacuation`, `heavy rains`, `disrupted transport`, etc. provide better comprehension when seen all together.

- **Event spreading:** As the event evolves, it propagates spatially to neighboring regions or specific geographical locations, which are determined by its inherent nature. For example, `Hurricane Katrina` spread over the gulf coast of the USA, whereas `WannaCry Ransomware`[10] spreads to many countries across continents but not to a few selected countries. However, in current scenario, event diffusion can occur via non-conventional channels over the Web.

- **Event sequencing:** The evolution of events involves a temporal sequence that is more likely to occur relative to all possible combinations. For example, the subsequent `flooding` after the `Hurricane Katrina`, and `missing persons` as a result. Here, these events are more likely to occur in a sequence such as `Hurricane Katrina` → `flooding` → `missing persons`.

- **Event interaction:** While on their trajectory, events interact with their surrounding entities and events, and cause other events to happen. For example, the merger of `tropical wave` with `Tropical Depression Ten`[11] initiated the formation of `Hurricane Katrina`. The sensitivity of an event towards certain entities and events, is highly influenced by their nature.

[9] https://en.wikipedia.org/wiki/Hurricane_Katrina
[10] https://en.wikipedia.org/wiki/WannaCry_ransomware_attack
[11] https://en.wikipedia.org/wiki/Tropical_Depression_Ten_(2005)

1.2.1 Event Impact Analytics

Societal events have an impact and consequences on the future. In such a scenario, the knowledge of how an event will spread in the temporal and the geographical space, carries utmost value. The impact gets reflected in varied forms depending on the event characteristics. For instance, a computer server failure because of high traffic as a result of some societal event, or catastrophic casualties caused by natural disasters, etc. The knowledge about the future unfolding/spread of an event, enables the respective authority(ies) to make proactive preparations for a better handling of the impact. The spread of an event can be easily imagined physically in spatial dimension, but provided the state of current information age, an event can propagate through other diffusion channels on the Web that carry high significance. For example, from an inter-cultural perspective, the diffusion of a societal event among foreign language communities.

Nowadays, societal happenings of disparate kind receive almost global diffusion because of the tremendous growth in coverage of news from virtually all over the world. The world-wide users of Facebook are reported to have just a 3.5 degree of separation[12], which implies that if an event is of significant interest to a particular user community, then it will easily find its way there. The range of diffusion varies greatly by the scale of the incident, for example a relatively small-scale event such as `Oscars 2017 mix-up` with a hyper-local context is less likely to have a global diffusion as compared to the `Brexit`[13] that concerns much broader section of public. The smaller incident might get limited to short term reactions in social media only, while the event concerning broader audience will have a far-reaching impact. Eventually, a societal relevant event might get reflected into a collaboratively curated encyclopedic platform such as Wikipedia. Following our expectation, the impact of a societal relevant event is not necessarily bound to a single community. For example the aforementioned `Brexit` has been translated to more than 60 languages in addition to the language of its origin, i.e., English. Interestingly, these comprised of around 50 predominantly spoken languages in the affected countries, which means majority of the European languages. Moreover, languages from affected countries are also few of the earliest ones to get translations.

However, the crucial question raised here is that how can one automatically assess the impact of societal relevant events with respect to their diffusion potential into foreign language communities? To answer this, we hypothesize that entities involved in an event carry semantic knowledge that provide strong contextual clues to successfully predict the diffusion of events in foreign language communities. In the following subsection, we discuss the importance of entities in analyzing the impact of events, and in effectively capturing semantics of Web contents.

1.2.2 Contextual Importance of Entities

As aforementioned, the evolution and trajectory of events are susceptible to their interaction with neighboring real world entities. In several examples of societal events such as `Hurricane Katrina`, `WannaCry Ransomware`, `Brexit`, etc., it is trivial to observe that the

[12] https://research.fb.com/three-and-a-half-degrees-of-separation/
[13] https://en.wikipedia.org/wiki/Brexit

kind of impact made by such events is strongly influenced by named entities they affected or involved. Here, for example, in case of the `Brexit`, geographical regions to which this event have affected and gained higher popularity, are strongly influenced by the associated entities (i.e., countries in `European Union`, and more). Bag-of-words is one of the most commonly used representation of documents for several prediction tasks, which basically represents a document by the words it contains, and computes their frequency. It does not use canonicalized named entities as well as ignores any entity-level information. Thus, it is difficult to capture the document semantics by using bag-of-words representation. This raises the question: how do we represent or process a document to capture its semantics and gain deeper understanding? In order to do so, a human interpretation is required that would facilitate achieving various higher level goals such as the prediction of event spread among certain communities or the classification of documents in fine-grained categories. While natural language processing (NLP) techniques are making good progress, a sophisticated understanding/interpretation of contents is still in its infancy. However, important clues can still be derived from the content's "inherent semantics". Our approach therefore aims to exploit entity-level analytics in order to capture a deeper interpretation of Web contents and perform various prediction tasks.

There are several benefits that come along with raising Web contents to the entity-level. One of these benefits is the viability of incorporating diverse knowledge from knowledge bases (KBs). Nowadays, there are several openly available large-scale knowledge sources that contain a wealth of information about named entities, such as DBpedia [Auer et al., 2007] or YAGO [Suchanek et al., 2007, Hoffart et al., 2013]. Furthermore, with the emergence of tools for interlinking text documents with Linked Open Data (LOD) like AIDA [Hoffart et al., 2011, Yosef et al., 2011] or DBpedia Spotlight [Mendes et al., 2011] there are efficient means to augment semantics to plain text by raising it to the entity-level. By doing so, we are able to exploit the structured knowledge stored in KBs and derive valuable background information about the involved entities, which can help in achieving better performance for the aforementioned high-level prediction tasks.

1.3 Approach and Contributions

The salient contributions of this thesis are three-fold. The first is event spread prediction. To accomplish that, we present a framework that harnesses entity-level semantics and predicts events diffusion in foreign language communities. The second is entity-level content classification, for which we propose a novel approach for the semantic representation of documents namely, "semantic fingerprinting". And the third is news virality assessment and visualization. An online implementation of this work is deployed on the Web as an interface for the public use.

Each of the three aforementioned tasks is inherently difficult because of their complex nature. It is not easy to predict the event spread as it requires semantic understanding of various event characteristics. Also, the alignment of Web contents to a fine-grained type hierarchy, demands capturing the subtle differences in nature of various types. For news virality analytics, one has to discover semantic connections and capture the social perception, which is also not straight forward. To address this, the conceptual approach

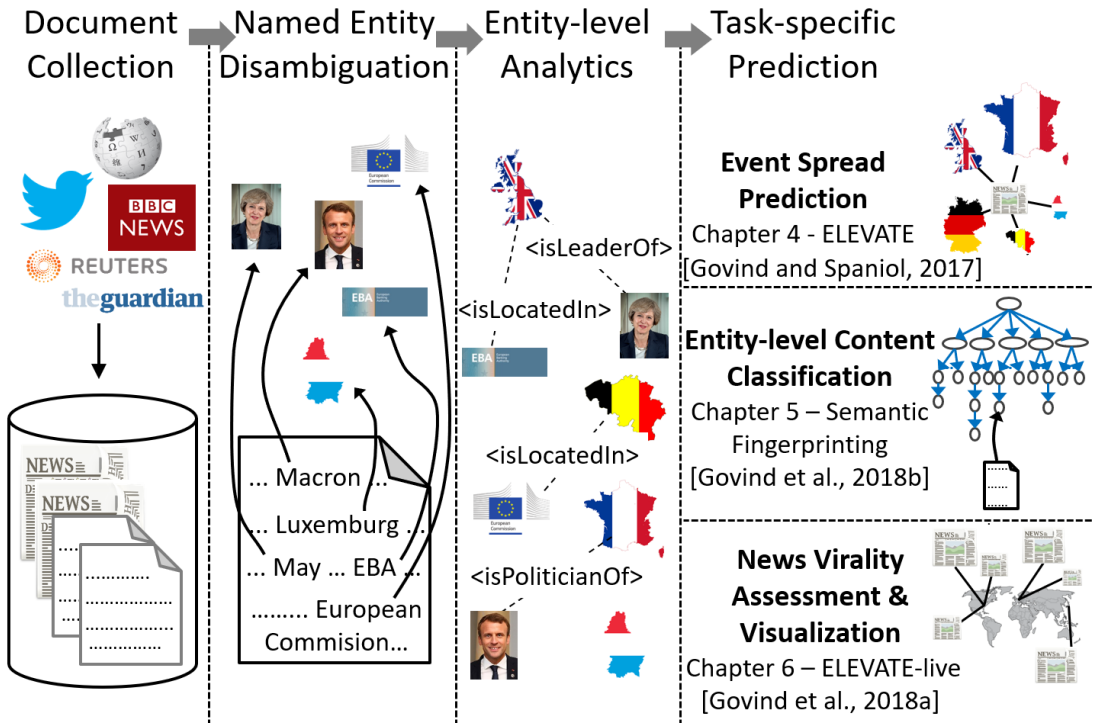


Figure 1.2: Overview of the conceptual approach based on entity-level analytics

behind our work has entities at its core. We raise the content of a document to the entity-level to better capture its semantics. In turn, it facilitates in gaining a deeper semantic understanding which can be utilized for a variety of prediction tasks. Figure 1.2 depicts an overview of the conceptual approach of the work done in the scope of this thesis. It depicts that the initial steps involve the extraction and disambiguation of named entities, followed by the entity-level analytics to utilize societal knowledge distilled and derived from KBs. Subsequently, the semantically augmented documents can be utilized by task-specific prediction algorithms to achieve disparate goals. We provide a brief overview of each of the aforementioned contributions in following subsections.

1.3.1 Event Spread Prediction

The event spread prediction task deals with prediction and analysis of the spread/diffusion of societal events via the information gathered from cyberspace. This is of high importance as it provides insights into the future evolution of daily happenings in the society. The early on information about the evolution of an event can be used for several purposes such as allocating resources appropriately to accommodate the impact/spread of events, etc. In order to investigate the diffusion of unscheduled events (such as natural disasters, conflicts, political actions, etc.), we introduce ELEVATE (Entity-LEVel AnalyTics for Event diffusion prediction) framework. The ELEVATE framework performs the semantic analysis of Web contents by utilizing entities contained and resources from LOD (YAGO knowledge base [Suchanek et al., 2007], here). In contrast to the existing approaches, our approach exploits the entity-level knowledge incorporated via knowledge bases. After the

systematic aggregation of entity-level knowledge, an event diffusion prediction model is trained that learns to predict the spread of events into foreign language communities. Our event diffusion prediction approach is driven by the following hypothesis:

> "*The spread of an event into foreign language communities depends on the named entities involved*". This can be broken down in the following aspects:

Hypothesis 1.1 The incorporation of semantics from LOD resources via entity-level analysis of Web contents, should allow to build a "semantic" prediction model.

Hypothesis 1.2 A machine learning model (suitably a multi-label classifier) assisted by entity-level analytics, can learn to identify typical diffusion patterns and predict spread for events.

Hypothesis 1.3 Based on named entities contained, an event spread on associated/involved countries should be observable.

1.3.2 Entity-level Content Classification

Content classification is the task of assigning appropriate classes/types/categories to Web contents, or documents in general. It has broad spectrum of applications, which vary by the number of output classes being targeted (e.g. sentiment analysis, type classification, etc.). In this work, we focus on the problem of aligning Web contents onto a fine-grained type hierarchy. In line with the core idea of this thesis, i.e., exploiting semantics via entities, we introduce "Semantic Fingerprinting" as a novel approach towards fine-grained entity-level content classification. The semantic fingerprinting approach captures inherent semantics of a Web content via the type information of contained entities with the help of YAGO knowledge base. To this end, we investigate the classification of Web contents to a fine-grained type hierarchy (>100 types) by harnessing entity-level semantics. As a vital part of this study, we also investigate the classification of events to around 20 fine-grained types. We formulate the key hypothesis behind this work as follows:

> "*A document can be characterized by the named entities it contains*". Therefore, we postulate he following:

Hypothesis 2.1 The semantics of a Web content can be captured via named entities.

Hypothesis 2.2 A concise and quality representation, i.e., "semantic fingerprint" of Web contents can be obtained by exploiting the entity-level type information.

Hypothesis 2.3 "Semantic Fingerprints" when backed by machine learning can provide a sophisticated fine-grained type classification model for Web contents.

1.3.3 News Virality Assessment and Visualization

News virality prediction is the task of automatically assessing the popularity of news articles at a time point in future. In this work, we present ELEVATE-live system, which is an extension of our event diffusion prediction framework ELEVATE [Govind and Spaniol, 2017]. The ELEVATE-live is a news virality analytics platform, which facilitates end users to explore and visualize the virality/relevance of Web news articles with respect to geographical locations (specifically, countries). In summary, the following notable hypothesis is explored as part of this work:

> "*Entity-level semantic analysis driven by the ELEVATE framework can reveal various characteristics regarding the virality of Web news articles*". In turn, we hypothesize the following:

Hypothesis 3.1 Non-trivial background information for a news article can be revealed by exploiting the named entities contained.

Hypothesis 3.2 Associated geographical regions can be explored by exploiting LOD.

Hypothesis 3.3 Interactive access mechanisms allow the inspection and visualization of the "virality" of news articles with respect to geographical locations concerned.

1.4 Publications in the Scope of the Thesis

Work in this thesis has been published at renowned international conferences. In particular, we have addressed conferences bridging the gap. ELEVATE framework for the event diffusion prediction has been published in the 9^{th} ACM Web Science Conference 2017 [Govind and Spaniol, 2017]. We have been awarded travel grant to present the ELEVATE framework. The Web Science Conference is one of the premier venue addressing the methods and findings to develop understanding of the Web. The conference aims to bring together the researchers working in multiple disciplines such as computer and information science, sociology, psychology, etc. The document understanding and representation based on "Semantic Fingerprinting" for the content classification is published in the International Conference on Web Engineering (ICWE) 2018 [Govind et al., 2018b]. The ELEVATE-live framework for the news virality assessment and visualization is also published in ICWE 2018 [Govind et al., 2018a]. The ICWE conference is a reputed venue for research on emerging challenges in the engineering of Web applications, social and cultural impact of the Web. Also, a scholarship was awarded by WSNET Web Science Summer School 2018 to present a poster [Govind, 2018] on the work done in this thesis.

Govind and Spaniol, M. (2017). ELEVATE: A Framework for Entity-level Event Diffusion Prediction into Foreign Language Communities. In Proceedings of the 9th International ACM Web Science Conference (WebSci'17), pages 111–120.

Govind, Alec, C., and Spaniol, M. (2018b). Semantic Fingerprinting: A Novel Method for Entity-Level Content Classification. In Proceedings of the 18th International Conference

on Web Engineering, ICWE 2018, Caceres, Spain, June 5-8, 2018, pages 279–287.

Govind, Alec, C., and Spaniol, M. (2018a). ELEVATE-Live: Assessment and Visualization of Online News Virality via Entity-Level Analytics. In Proceedings of the 18th International Conference on Web Engineering, ICWE 2018, Caceres, Spain, June 5-8, 2018, pages 482–486.

Govind (2018). Entity-level Event Impact Analytics. WSTNET Web Science Summer School, Hannover, Germany, July 30 - Aug 4, 2018.

1.5 Structure of the Thesis

The overall structure of the thesis is organized as follows. In Chapter 2, we provide a detailed introduction to fundamentals and the technical background necessary to understand the work conducted as part of this thesis. We explain the idea of Linked Open Data (LOD) and knowledge bases (KBs). In addition, we present fundamental concepts of machine learning models and discuss in detail about the ones that are employed in our work. We also discuss some of the widely employed evaluation methods that are well suited for tasks investigated here.

Subsequently, in Chapter 3, we discuss the prior work and survey the related research in detail. We summarize and compare a variety of social event analytics studies ranging from event detection to impact analysis, up to the virality prediction and analysis. Then we survey disparate works based on knowledge base models and entity-level analytics. We conclude this chapter by discussing recent studies on the type classification task. We address both the entity type classification as well as the document type classification but focus more on the document level works as it is more relevant to this thesis.

Furthermore, in Chapter 4, we present our work on the event diffusion in foreign language communities. Here, we introduce the ELEVATE (Entity-LEVel AnalyTics for Event diffusion prediction) framework. We report results of our extensive experiments conducted on the spread prediction task over around hundred languages. We consider several state-of-the-art baselines and multiple variations of our model. Moreover, we study the evolution of various performance measures by conducting experiments at three time points after the beginning of an event. We introduce a machine learning model that learns typical spread patterns and improve the performance. Finally, we summarize our overall findings from experiments on the event spread prediction task.

Subsequently, in Chapter 5, we address the task of content classification into fine-grained types. Discovering appropriate fine-grained type labels for Web contents and events, is a fundamental task in organizing information as it acts as a seed to many higher level tasks. To this end, we propose “Semantic Fingerprinting”, that provides a concise and effective semantic representation of documents based on entities contained. We employ the proposed approach to align Web contents onto a fine-grained hierarchy of 105 types that is comprised of 20 most frequently used subtypes for each of the 5 top-level types (i.e., event, organization, location, artifact, and person). We report our results on a dataset from Wikipedia and perform comparison to several state-of-the-art competitors.

In Chapter 6, we investigate the task of news virality analytics from its geographical perspective. We introduce ELEVATE-live, which is an extension of ELEVATE framework to assess the virality of online news articles. Likewise the overall theme of the thesis, the ELEVATE-live is also purely based on entity-level analytics. The ELEVATE-live provides the mechanism to assess as well as visualize a news article's relevance with respect to geographical locations. For that purpose, an online interface has been built and deployed as part of ELEVATE-live, which is available freely on the Web for public use. As part of this work, we monitor a variety of reputed online news agencies, and enable the end user to check the virality of news in real time.

Finally, in Chapter 7, we conclude the overall work performed and provide an outlook for the prospects of future research opportunities that have been opened up by this thesis. Although, we chapter-wise summarize our findings, in this chapter we look upon the work in a more holistic manner and provide a broader perspective.

Chapter 2

Foundations and Technical Background

In this chapter, we introduce the technical background required to understand the work done as part of this thesis. We start with the explanation of knowledge bases and linked open data. Subsequently, we introduce the need and challenges of entity extraction and disambiguation, followed by machine learning techniques for classification problems. We explain several link prediction methods as they are utilized in our work. At last, we provide an detailed overview of relevant evaluation measures.

2.1 Knowledge Bases and Linked Open Data

A knowledge base (KB) is the representation of common knowledge accessible in a machine readable format that enables machines to comprehend semantic data. The term knowledge

baseis referred to RDF (resource description framework) data sets published based on a set of linked data principles proposed by W3C (World Wide Web Consortium) as a standard for the knowledge representation. The idea of explicit representing the semantics of data was introduced by Tim Berners-Lee et al. [Berners-Lee et al., 2001] by coining the term "Semantic Web". These common principles are receiving a growing acceptance since then. Now a days, there exist many RDF datasets published by public and private organizations [Lütkebohle, 2008]. Linked data present the set of practices to publish and connect structured data on the Web, and linked data that are available with open access are termed as Linked Open Data (LOD) [Bizer et al., 2011].

2.1.1 Resource Description Framework

The resource description framework (RDF) specifies a data model proposed by W3C as a standard for representing information on the Web [W3C et al., 2014]. The RDF standard guides the construction of a graph which is formed by subject-predicate-object triples. The elements of these triples are resources, predicates, and datatyped literals which are used to express descriptions of resources. RDF datasets are comprised of a default graph, and zero or more named graphs.

A resource in the RDF model refers to an entity or a concept. Let us denote the set of all resources in the knowledge base with R and literals by L. A literal is used to represent the data values of various types such as a real number, string, date, time etc. The other important elements of the RDF model are predicates, P, which serve to represent a relation among resources, or a resource and a literal. Now, a KB can be defined as the projection between these building blocks of the RDF data model. To this end, a knowledge base can be represented as following: $K = R \times P \times (R \cup L)$. Alternatively, a KB can be seen as simply a set of triples of the form $t =< s, p, o >$ where $s \in R$, $p \in P$, and object o can represent a resource or a literal, thus, $o \in R \cup L$. Table 2.1 provides an example of RDF graph for the resource `Albert Einstein`.

Subject	Predicate	Object
dbr:Albert_Einstein	rdf:type	dbo:Scientist
dbr:Albert_Einstein	dbo:award	dbr:Nobel_Prize_in_Physics
dbr:Albert_Einstein	dbo:knownFor	dbr:Special_relativity
dbr:Albert_Einstein	dbo:knownFor	dbr:Mass-energy_equivalence
dbr:Albert_Einstein	dbo:birthPlace	dbr:Ulm
dbr:Albert_Einstein	dbo:citizenship	dbr:Switzerland
dbr:Albert_Einstein	dbo:citizenship	dbr:Statelessness
dbr:Albert_Einstein	dbo:field	dbr:Physics

Table 2.1: An RDF graph example - RDF triples for resource `Albert_Einstein` in DBpedia; `dbr` and `dbo`, stand for DBpedia resource, and ontology/schema, respectively

2.1.2 Resource Description Framework Schema

The categorization of RDF resources among various classes is of great importance in representation of the real world knowledge. Resource Description Framework Schema (RDFS) provides the set of classes for the description of ontologies [W3C et al., 2014]. RDFS enables the RDF data model in categorizing resources to an intended structure. As depicted in Table 2.2, a class in RDFS is defined by a set of triples. To this end, a class c is assigned to a resource r by utilizing the predicate `rdf:type` such as $< r, \texttt{rdf:type}, c >$.

Subject	Predicate	Object
dbo:Scientist	rdf:type	owl:Class
dbo:Scientist	rdfs:subClassOf	dbo:Person
dbo:Person	rdfs:subClassOf	dbo:Agent
dbo:Agent	rdfs:subClassOf	owl:Thing
dbo:Agent	owl:disjointWith	dbo:Place

Table 2.2: An RDFS definition example - RDF triples specifying a segment of DBpedia ontology/schema

RDFS enables the introduction of a well-defined hierarchy among classes assigned to resources in a KB. For example, the resource `Albert_Einstein` is an instance of the class `Scientist`, as listed in Table 2.1. By utilizing the predicate `rdfs:subClassof` in the RDFS definition, it can be deduced that `Albert_Einstein` belongs to the `Person` class as well. To this end, RDFS provides a mechanism to define the structure and the appropriate hierarchy among the resource classes.

2.1.3 Large Scale Knowledge Bases

Since the publication of the Semantic Web standards by W3C, the idea of Linked Open Data (LOD) has grown popular. Nowadays, the number of openly available RDF data sets are in thousands and continuously growing [Lütkebohle, 2008]. Some of the popular LOD examples are DBpedia [Bizer et al., 2009], YAGO [Suchanek et al., 2007], Freebase [Bollacker et al., 2008], Wikidata[14], etc. Traditionally, the construction of KBs has been done by employing human experts (such as the construction of WordNet [Miller, 1995]). The creation and the maintenance of a large scale KB is very expensive if we rely on manual work of a few experts. Recently, with the growing presence of the Web in our society and advancements in crowd sourcing platforms, it has become relatively easier to employ the wisdom of crowd. Wikidata is such an example of LOD cloud data set, which is curated by volunteers coming from diverse geographical regions and backgrounds. Another commonly used technique to populate a real world KB is by automatically extracting and distilling facts from semi-structured data (e.g., Wikipedia). DBpedia and YAGO are examples of such KBs, which are constructed using automatic fact-extraction techniques from Wikipedia. These KBs are comprised of RDF triples about several million entities, and are available in multiple languages (10 major languages for YAGO, whereas DBpedia provides localized datasets for more than 100 languages).

[14] https://www.wikidata.org/

2.2 Named Entity Recognition and Disambiguation

Named Entity Recognition and Disambiguation (NERD) is the task of recognizing (cf. Section 2.2.1) the mention of entities in raw text and further aligning (cf. Section 2.2.2) them to the correct instance in a KB (e.g., DBpedia, YAGO, or Wikidata). We provide a detailed overview of both subtasks in the following subsections. There are broadly two classes of approaches for NERD: 1) Independent models for the recognition and the disambiguation tasks, 2) End-to-End models. The first class of approaches treats the recognition and the disambiguation tasks disjointly, and pipeline the separately built models for individual tasks. Thus, the disambiguation model can be fed with the entity mentions extracted from certain recognition model or ground truth of entity mentions itself if available. In contrary to the first approach, the second approach directly takes the raw text as input and outputs the disambiguated entities. It operates jointly on the extraction of named entities and their disambiguation to the correct entry in a given KB.

2.2.1 Entity Recognition

Named Entity Recognition and Classification (NERC) is one of the key information extraction tasks which seeks to identify the mention of entities in the text such as organizations, persons, locations, etc. [Nadeau and Sekine, 2007]. It operates typically in two phases: entity detection and entity classification. The entity detection step deals with extraction of named entity mentions in the text, and the entity classification step aligns an entity mention to a certain class. Some of the commonly used named entity recognition and classification systems include Stanford NER [Finkel et al., 2005], NLTK [Bird et al., 2009], spaCy[15], etc. A NERC system does not work upon resolving any ambiguity among the named entities extracted.

2.2.2 Entity Disambiguation

Named entity disambiguation (NED) deals with the task of resolving an entity mention in the text to a canonical entity in some given knowledge base. The challenge of NED task lies in the inherent ambiguity among the entity mentions. For instance, there can be multiple named entities being referred using the same surface form in the text (e.g., "Hathaway" can refer to the company `Berkshire Hathaway Inc.` or the actress `Anne Hathaway`). Many of the large scale disambiguation systems are built around resources from the Wikipedia ecosystem. The link anchors in Wikipedia pages are commonly used by multiple systems for collecting variations in entity mentions. Moreover, Wikipedia pages serve broadly as the ground truth for real word entities. Some of the popular entity disambiguation systems are DBpedia Spotlight [Mendes et al., 2011], AIDA [Hoffart et al., 2011], TagMe [Ferragina and Scaiella, 2010], etc. Moreover, there have been recent work that employ deep learning techniques and distributional semantics. One of these work is by [Moreno et al., 2017], where the model jointly learns the embeddings for words in text and named entities in a knowledge base in order to perform accurate disambiguations.

[15] https://github.com/explosion/spaCy

In the following, we present an overview of some of the most widely used named entity disambiguation systems.

DBpedia Spotlight

DBpedia Spotlight was developed to link entity mentions in a given text to the DBpedia resources [Mendes et al., 2011]. The approach behind DBpedia Spotlight operates in four stages. The first stage is the spotting stage, which identifies the phrases that might be a mention of potential DBpedia resource. In the subsequent stage, the candidate selection is performed. Here, potential DBpedia resources for each of the mentions are assembled. The next stage is the disambiguation stage where the best amongst the candidate resources are selected for each spotted mention with the help of context around it. The context of the surface form and DBpedia resources are represented by a tf-icf (term frequency-inverse candidate frequency) vector in a multidimensional space of words. A weight metrics namely, inverse candidate frequency (ICF) is employed as shown in Equation 2.1. Here, R is set of candidate resources for surface form s, and $n(w)$ gives the number of resources related to word w in set R. The objective is to rank the correct resource for a mention at highest with respect to its similarity to the context of mention. Finally, the annotation process can be configured to the best of user needs by using the parameters such as target resource set, prominence of a resource, topical relevance, disambiguation confidence, etc.

$$ICF(w) = log\frac{|R_s|}{n(w)} \tag{2.1}$$

AIDA

Accurate Online Disambiguation of Named Entities (AIDA) is a NED system that formulates the disambiguation problem as a dense sub-graph detection problem by proposing a weighted mention-entity graph [Hoffart et al., 2011, Yosef et al., 2011]. AIDA employs Stanford NER for extracting mentions from the input text and the knowledge base YAGO as the target entity set. The mention-entity graph is a weighted undirected graph where nodes are mentions and entities. A weight is assigned to each edge between a mention and an entity based on a similarity measure, or a combination of similarity and popularity measures (i.e. prior). On the other hand, an edge between entities can be assigned a weight based on the coherence (Equation 2.2) among the entities, or type distance, or some combination of the two. AIDA incorporates the mention-entity graph as can be seen in Figure 2.1. Equation 2.3 was proposed by [Milne and Witten, 2008] to measure the relatedness between two entities e_1 and e_2 based on the shared inlinks on Wikipedia, and is used by AIDA for the computation of coherence. Here, E_i represents the set of inlinks to entity e_i and E is the set of all entities. AIDA provides the end user options of three disambiguation schemes based on the aforementioned measures, i.e., *prior*, *prior* + *similarity*, and *prior* + *similarity* + *coherence*.

$$coherence(e_1, e_2) = \left\{ \begin{array}{cc} 1 - relatedness(e_1, e_2) & \text{if } >0 \\ 0 & \text{otherwise} \end{array} \right\} \tag{2.2}$$

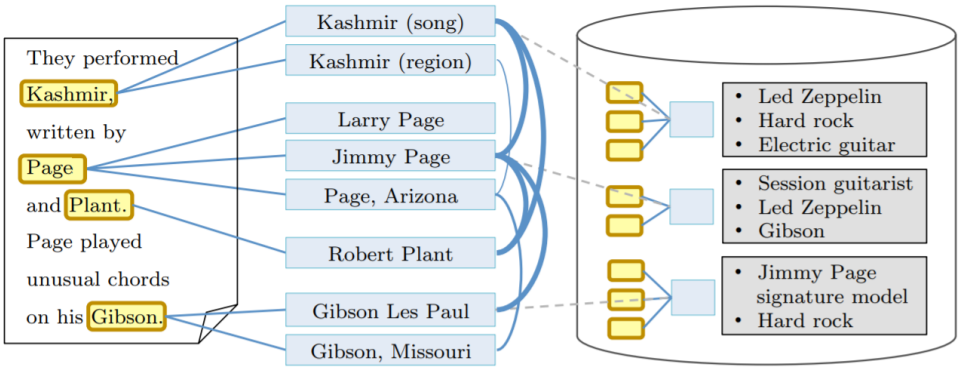


Figure 2.1: An example of the mention-entity graph in AIDA [Hoffart et al., 2011]

$$relatedness(e_1, e_2) = \frac{log(max(|E_1|, |E_2|)) - log(|E_1 \cap E_2|)}{log(|E|) - log(min(|E_1|, |E_2|))} \tag{2.3}$$

TagMe

A system designed to augment the input plain text with hyperlinks to Wikipedia entities is TagMe [Ferragina and Scaiella, 2010]. TagMe is developed while keeping in mind to optimize its working on short text such as tweets, news articles, etc. It is a quite handy feature as there is a vast number of Web contents of relatively small size. Like the aforementioned disambiguation systems, TagMe also utilizes anchor texts from Wikipedia. A target sense (i.e., Wikipedia page) p_a for a mention a is scored based on the voting from the other mentions $b \in A_t - a$ in the input text, where A_t is the set of all mentions in the text. If a mention b is unambiguous, then its vote for p_a is equal to the relatedness between p_a and p_b. On the other hand, when b is ambiguous then the most related sense to p_a will dominate its vote. TagMe also uses the relatedness score proposed by [Milne and Witten, 2008] as previously seen in Equation 2.3.

2.3 Supervised Learning and Classification Methods

Supervised learning is a task in machine learning that aims to learn an input to output mapping function based on a given set of input-output example pairs. This learned mapping function can be further used to annotate unseen examples. A general formulation of the supervised learning problem can be devised as follows:

Given a training set $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ of size N such that (x_i, y_i) is the i^{th} example pair where x_i represents the feature vector and y_i represents the corresponding output (i.e. some real number or class label depending on the nature of problem). It becomes a regression problem when the output is a real number, whereas when the output is a class label, the problem belongs to classification. Now the objective of the machine learning algorithm is to learn a function $f : X \rightarrow Y$ from the training data whose domain X is the input space and range Y is the output space, and function f belongs to the hypothesis space of all feasible hypothesis functions. The training procedure aims to minimize a loss/risk function in order to fit a model on the training set.

In machine learning, classification is the task of discovering the output class/category label $c \in C$ for an unseen instance x based on the learned inherent patterns in a given training data where output labels are known. In a regression problem the output is a real number, on the contrary, the classification problem seeks the prediction of categorical values (e.g., type label for Web contents). In view of the work done as part of this thesis, supervised learning and especially the classification is more suitable and employed extensively. We provide a detailed explanation of some of the widely used classification algorithms in the following subsections.

2.3.1 Naive Bayes

Naive Bayes (NB) classifier belongs to the family of probabilistic classifiers, and is based on Bayes' theorem. Naive Bayes classifiers assume the conditional independence among the input features. NB classifier is considered as one of the default baselines for several natural language processing tasks, such as the text categorization. For such tasks, the tf-idf (term frequency-inverse document frequency) features are quite commonly used. The term frequency captures the weight of a term in the document, whereas the inverse document frequency quantifies the specificity of a term. We define the set of all classes by C and the input feature vector for an example by x. The most likely class or maximum a posteriori (MAP) can be formulated as given in Equation 2.4.

$$C_{MAP} = \arg\max_{c \in C} P(c|x) \tag{2.4}$$

By using Bayes' theorem and dropping the denominator (as it remains constant among all classes), the most likely class can be computed using Equation 2.5.

$$C_{MAP} = \arg\max_{c \in C} P(x|c) * P(c) \tag{2.5}$$

As the NB classifier considers conditional independence among the input features, i.e., the feature vector x's entries $x[1], x[2], ..., x[F]$ are independent, where F is features count. Finally the most likely class prediction can be done using the formula as shown in Equation 2.6. To this end, we can assign the class label $\hat{y} = C_{MAP}$ to the test instance x.

$$C_{MAP} = \arg\max_{c \in C} P(c) \prod_{i=1}^{F} P(x[i]|c) \tag{2.6}$$

The training of Naive Bayes classifiers can be done by evaluating only closed-form expressions (i.e., likelihoods, and priors). The NB training process is rather inexpensive as compared to the iterative approximation methods used for many other types of classifiers such as logistic regression, support vector machines, etc. The multinomial formulation of Naive Bayes is suitable for problems such as document classification. In practice, there can arise several problems to a NB classifier such as zero probabilities, and underflowing probability values. The regularization techniques such as Laplace smoothing can be employed to deal with zero probabilities. To tackle the underflow of very small probability values, the NB classifier can be expressed in log-space. Now, small probability values are not being multiplied but rather added, thus, avoiding the underflow. A multinomial NB classifier becomes a linear classifier when expressed in the log-space.

2.3.2 Random Forests

Random Forest (RF) or random decision trees, are an ensemble method that employs a collection of decision trees to learn a better hypothesis over the training data [Breiman, 2001]. The RF classifier combines the bagging ensemble approach with random decision trees to achieve a high predictive performance and improved stability. As a multitude of decision trees participate in deciding the output label for a test instance, the output is the average of values predicted by individual trees in case of regression, and the majority voting technique is utilized to get the output label in case of classification. An individual decision tree model is highly prone to the problem of overfitting/high variance and thus, performs poorly on unseen data due to lack of generalization. The RF algorithm tackles the problem of overfitting by building a collection of uncorrelated decision trees. The basic idea of the RF training process is to re-sample the training set over and over, and for each sample it trains a new decision tree. Different decision trees may overfit the training set in a different way, but via averaging/majority voting those differences are averaged out. The training of individual decision trees can be done using algorithms such as ID3, C4.5 [Quinlan, 1993, Michalski et al., 2013], etc., which create the tree by splitting the branches based on the feature with highest information gain. The Equation 2.7 formulates the information gain provided by a feature f on a set T of training examples where $T(v)$ represents the set of examples with the value v of feature f. The *entropy* function measures the impurity in an arbitrary collection of examples and *values* function gives all present values of a feature in the given collection of examples.

$$Gain(T, f) = Entropy(T) - \sum_{v \in values(f)} \frac{|T(v)|}{|T|} Entropy(T(v)) \tag{2.7}$$

The following two principles govern the construction of an ensemble of uncorrelated decision trees in the RF algorithm.

Bagging

Bagging or bootstrap aggregating produces multiple versions of the original training set with the help of random sampling, and in a subsequent step multiple models are built using these individual training sets to form an ensemble of models. Given the number of decision trees s to build as part of the RF model and a training set $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ of size N, the bagging algorithm generates s new training sets, i.e., one for each of the individual decision trees. Each of the new training sets $T_i \in \{T_1, T_2, \ldots, T_s\}$ is generated by sampling uniformly with replacement from the original training set T for N times. As sampling with replacement is used, these training sets can contain multiple copies of an example from the original training set. Having a different training set for each of the decision trees, helps in building decision trees with lower correlation among them, and thus, the ensemble of trees generalizes better on the original training data set.

Random Subspace Method

Tin Kam Ho proposed random subspace method to reduce the correlation between learned decision trees [Ho, 1998]. It is also commonly known by terms such as attribute bagging or

feature bagging. The random subspace method differs from the bagging algorithm in the aspect that it operates on the feature set instead of the training set. It attempts to reduce the correlation between the decision trees by training them on a random sample of features instead of the entire feature set. For each of the individual estimators (i.e., decision tree, here), the random subspace method selects a different subset of input features randomly. The motive behind using only a subset of features, is to avoid the reliance on a small number of highly predictive/discriminative features as those features might not be as predictive for unseen test data as they are for the training data.

2.3.3 Support Vector Machines

Support vector machine (SVM) is a supervised learning algorithm that aims to construct a maximum-margin hyperplane to separate the examples by mapping them to high-dimensional space [Cortes and Vapnik, 1995]. SVM is a kernel-based method and has sparse solutions, which means that the predictions for unseen data depend only on the kernel function evaluated at a subset of data points in the training set. This subset of training data points are also commonly known as support vectors. As previously described in Section 2.3, T is a training set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ of size N such that (x_i, y_i) is the i^{th} example pair where x_i represents the feature vector and y_i represents the corresponding output label. For a binary classification problem, $y_i \in \{-1, 1\}$. Thus, a two class classifier can be formulated using the Equation 2.8 where sign of function $f(x)$ gives the output binary label for some test input example x. Here w and b are explicit parameters, and $\phi(x)$ denotes a feature-space transformation.

$$f(x) = w^T \phi(x) + b \tag{2.8}$$

Figure 2.2 depicts the illustration of maximum-margin hyperplane and margin of an SVM trained for a binary classification problem. The data points from training data, which lie on the margin are called support vectors.

If the training data is linearly separable, all the data points in training set should satisfy Equation 2.9, which is a canonical representation of the decision hyperplane.

$$y_i(w^T \phi(x_i) + b) \geq 1, i = 1, \ldots, N. \tag{2.9}$$

To generalize and separate the two classes in the training set, we need a decision hyperplane with maximum possible margin. Consequently, we need to maximize the distance between margin boundaries and the hyperplane (i.e., $||w||^{-1}$). Thus, the problem of finding the maximum margin hyperplane can be formulated as an optimization problem as show in Equation 2.10 subjected to constraints given in Equation 2.9.

$$\arg\min_{w,b} \frac{1}{2}||w||^2 \tag{2.10}$$

To solve the constrained optimization problem of Equation 2.10, we can transform the original problem with the help of Lagrange multipliers $\alpha_i \geq 0$ as given in Equation 2.11.

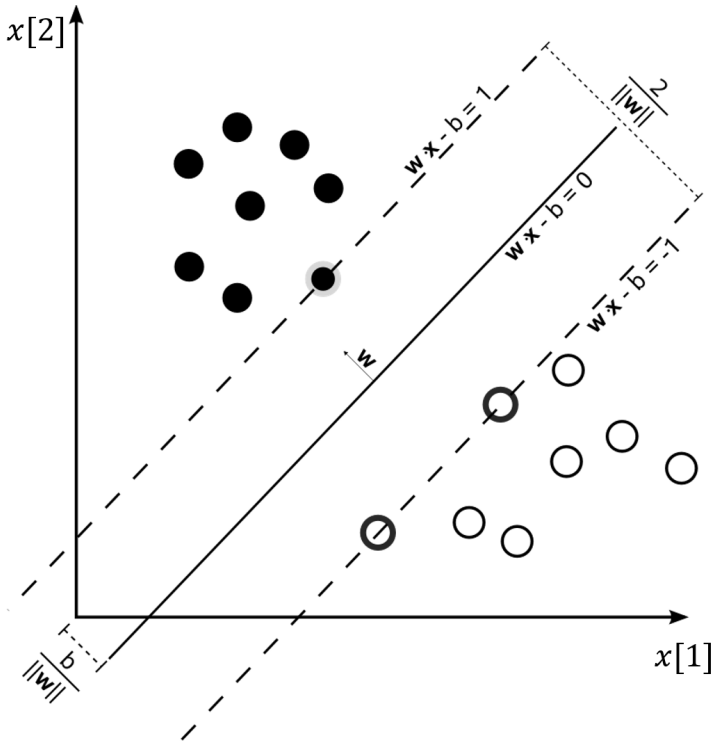


Figure 2.2: SVM hyperplane with margin for a binary classification problem

To this end, we need to solve the following optimization problem with respect to w, b and α to obtain the maximum margin hyperplane.

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{N} \alpha_i y_i (w^T \phi(x_i) + b) - 1 \tag{2.11}$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)^T$. Taking derivative with respect to w and b and setting them equal to zero gives us the following conditions.

$$w = \sum_{i=1}^{N} \alpha_i y_i \phi(x_i) \tag{2.12}$$

$$0 = \sum_{i=1}^{N} \alpha_i y_i \tag{2.13}$$

Now, parameters w and b can be eliminated from Equation 2.11 and a dual of the problem with respect to Lagrangian multipliers α_i has to be solved. In order to classify a test example x with the trained SVM model, we need to evaluate the sign of $f(x)$ as formulated in Equation 2.14, which is obtained by substituting the value of w from Equation 2.14 in Equation 2.8.

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b \tag{2.14}$$

Here the kernel function is defined by $K(x, x_i) = \phi(x)^T \phi(x_i)$. The concept of formulating a kernel as an inner product of the feature space is known as kernel trick or kernel substitution. The kernel trick opens many interesting possibilities for the transformation of feature space without actually transforming inputs individually. The main idea behind this is that as the SVM algorithm requires inputs in the form of inner product only, the inner product can be replaced by any other choice of kernel function. So far, it was assumed that the classes in the training set are linearly separable but in real world settings it is not always the case. This implies that the SVM classifier will struggle to generalize over the training data points for such cases. The SVM algorithm provides an enhancement to achieve a better performance by allowing some of the data points to be wrongly classified. To do this, *slack variables* ζ_i where $i = 1, \ldots, N$, are introduced for each of the data point in the training set. The goal behind introducing slack variables is to softly penalize the data points that exist on the wrong side of boundary of the margin.

2.3.4 Multi-label Classification

Multi-label classification involves the learning of multiple labels from a set of given examples, in contrast to the before mentioned single-label classification where each example is associated with a single label. Multi-label classification has been increasingly becoming a need for many real world modern problems such as protein function categorization, music genre classification, scene classification, etc. [Tsoumakas and Katakis, 2006]. There are two ways of approaching a multi-label classification problem namely by problem transformation or problem adaptation. The problem transformation methods transform the problem in binary classification whereas in contrast adaptation methods deal the problem in full form which involves more complexity. We use one-against-all problem transformation [Rifkin and Klautau, 2004] as it is considered suitable for the large scale classification [Tang et al., 2009]. The one-against-all problem transformation method transform the multi-label classification into a set of binary classification problem instances, where an individual classifier is trained for each of the possible output labels.

A test instance can have more than one output labels $c \in C$ annotated where C is set of all possible class labels. Let $x \in \mathbb{R}^F$ represents the feature vector of size F for some input test instance. Here, we have to compute the vector $f(x) \in \mathbb{R}^{|C|}$ where $f_i(x)$ is responsible for determining the membership of x in a certain target class $i \in C$. To this end, one binary classifier is trained for each of the class labels. In total, there will be a set of $|C|$ classifiers that will decide the output class labels of a test instance. This means that the goal of individual binary classifiers is to learn to separate one designated class label from all other class labels. The training set T is a collection of all given examples (i.e., feature vectors) such that $T \in \mathbb{R}^{N \times F}$, and corresponding set of class labels represented by $Y \in \{0, 1\}^{N \times |C|}$. Therefore, a decision function f has to be built for each class by transforming the training set such that the instances belonging to the concerned class are considered as positive examples, whereas all the others will represent negative examples. When an SVM classifier is employed, the decision function $f_i(x)$ can be formulated as given in equation 2.15 for a class with label i with respect to some test instance x. It should be noted that the classifier will be trained on a binary training set of instance-label pairs $(x_j, y_j), j = 1 \ldots N$ where $x_j \in \mathbb{R}^F$ and $y_j \in \{-1, 1\}$. Here, b_i is a bias parameter for

the class i. $K(x_j, x)$ and $\phi(x)$ represent kernel function and feature mapping as discussed in detail in the previous subsection.

$$f_i(x) = sign\Big[\sum_{j=1}^{N} \alpha_j y_j K(x_j, x) + b_i\Big], \alpha_j \geq 0 \tag{2.15}$$

2.4 Complex Networks and Link Prediction

Complex networks naturally appear in a lot of real world problems. Future link prediction has been applied to solve variety of tasks in disparate fields where the underlying data is in the form of a network [Wang et al., 2015]. Particularly, it has been used to analyze social networks such as for collaboration in co-authorship networks [Liben-Nowell and Kleinberg, 2007], for detection of links in terrorists association networks [Clauset Aaron et al., 2008], and for recommendation in the social networks. More recent work involve suggesting who to follow and why in social networks [Barbieri et al., 2014], and link prediction in coupled networks [Dong et al., 2015]. Link prediction methods are broadly categorized in three types namely, similarity based methods, maximum likelihood methods and probabilistic methods [Lü and Zhou, 2011]. Our decision to employ the local similarity based methods is driven by their effectiveness [Liben-Nowell and Kleinberg, 2007] and simplicity in computation. The methods based on node neighborhoods include common neighbors, Jaccard's coefficient [Jaccard, 1901], Adamic/Adar index [Adamic and Adar, 2003] and preferential attachment [Kim and Leskovec, 2011]. These methods use the local information of a node in order to predict its likelihood of forming a link with other nodes in the graph. To this end, we discuss several of these similarity indices in the following.

- **Common Neighbors:** is formally defined by Equation 2.16 where x and y are nodes in the graph, and Γ represents the neighborhood of a node. Thereby, the common neighbors (CN) of a node x to counterpart(s) y are the nodes which are in both $\Gamma(x)$ and $\Gamma(y)$.

$$CN(x, y) = |\Gamma(x) \cap \Gamma(y)| \tag{2.16}$$

- **Adamic/Adar Index:** is formulated by Equation 2.17. The Adamic/Adar (AA) is a weighted reformulation of the common neighbors and has shown improvements over the previous approach [Liben-Nowell and Kleinberg, 2007].

$$AA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log(|\Gamma(z)|)} \tag{2.17}$$

- **Jaccard's Coefficient:** is yet another similarity based link prediction measure, which is formally defined by Equation 2.18. The common neighbors coefficient struggles to capture the similarity when a node has too many neighbors and thus, ends up having many common neighbors with almost every other node. Jaccard's

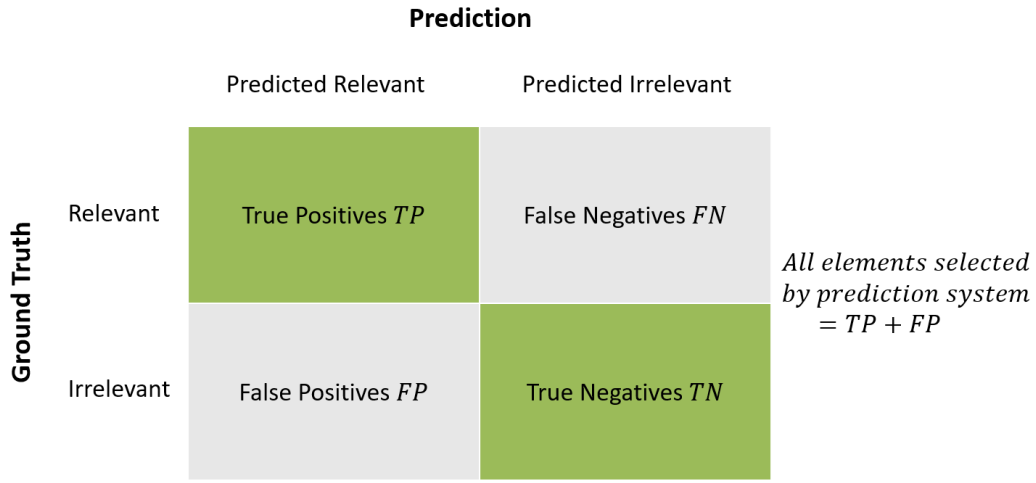


Figure 2.3: Different groups of elements while evaluating a test set

coefficient (JC) remedies the problem by dividing the number of common neighbors by the size of combined neighbors of the two nodes.

$$JC(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{2.18}$$

- **Preferential Attachment:** aims to quantify the notion of "rich gets richer". Preferential attachment (PA) is mathematically formulated in Equation 2.19. It is basically the product of the number of neighbors the two nodes have, which implies that if two nodes have a lot of neighbors then they will tend to have even more neighbors, and eventually they both will be neighbors of each other.

$$PA(x,y) = |\Gamma(x)| * |\Gamma(y)| \tag{2.19}$$

2.5 Evaluation

In this section, we introduce various performance evaluation metrics that are suitable for the classification tasks in this thesis. A classifier's goal is to retrieve/select all the elements in the test set that are relevant. In order to define any of the below mentioned evaluation measures, it is mandatory to first define the four groups of examples/elements that arise while evaluating a test set namely, true positives, false positives, true negatives, and false negatives. True positives (TP) are the elements in test set that are selected by the classifier and are relevant. False positives (FP) are the set of elements which are selected by the classifier but are not relevant. False negatives (FN) are the relevant elements that are not in the set of elements selected by the classifier. True negatives (TN) are the elements that are neither selected nor relevant. Figure 2.3 conceptually depicts the aforementioned sets of elements visually. Now, we are ready to define the standard evaluation measures as follows:

- **Precision** measures the correct instances out of all the selected ones by the classifier. It is formulated in Equation 2.20.

$$precision = \frac{TP}{TP + FP} \tag{2.20}$$

- **Recall** computes the fraction of relevant elements retrieved by the classifier over all the relevant elements as formulated in Equation 2.21.

$$recall = \frac{TP}{TP + FN} \tag{2.21}$$

- **F-Score** combines the precision with recall to measure performance of the classifier over the test set. A general F-score with a positive real parameter β is given by Equation 2.22. The F1 measure is more prevalent in practice and is defined as the harmonic mean of precision and recall as formulated in Equation 2.23.

$$F_\beta = (1 + \beta^2) * \frac{precision * recall}{\beta^2 * precision + recall} \tag{2.22}$$

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{2.23}$$

Micro/Macro-Averaged Scores

As discussed before, there are various classification tasks for which there can exist more than one output labels for individual test examples. To tackle such a scenario, macro and micro averaged scores are introduced. Let us assume that there are N_{test} number of examples in our test set. Also, let us denote the true positives for the i^{th} example by TP_i, and similarly the true negatives by TN_i, and so on. Now, micro-average scores are formulated using Equation 2.24 and 2.25. On the other hand, macro-average scores are basically the average of individual scores for each of the test examples as given in Equation 2.26 and 2.27. The macro and micro averaged F1 is computed by substituting the respective averaged precision and recall in Equation 2.23.

$$micro\text{-}precision = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FP_i} \tag{2.24}$$

$$micro\text{-}recall = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FN_i} \tag{2.25}$$

$$macro\text{-}precision = \frac{\sum_i precision_i}{N_{test}} \tag{2.26}$$

$$macro\text{-}recall = \frac{\sum_i recall_i}{N_{test}} \tag{2.27}$$

Chapter 3

Related Work

In this chapter, we provide a detailed discussion on prior and related work to our research. The state-of-the-art and related studies with reference to our work can be subdivided into thematic clusters as described in the following sections.

3.1 Societal Event Analytics

Societal event analytics deals with the detection, prediction and analysis of societal events from the Web data. We provide an overview of these tasks in the following subsections. Related to societal event analytics is culturomics [Suchanek and Preda, 2014], which studies the cultural trends and mass behavior in digitized text such as news reports.

3.1.1 Event Detection

With the widespread accessibility of the Web in our society, the data originating from societal events have grown exponentially. It has become intricate to give human attention to all of this available unstructured data in the current scenario. The task of event detection answers the questions such as ”What happened?” and ”What is new?”, and clusters the information on the Web by societal events.

Many studies have modeled an event as the collection of entities that are strongly connected with each other and have a high level of activity. The measure and notion of the connectedness, and the activity level may vary from one application domain to another. Related problems to event detection are community detection and anomaly detection, which are also widely formulated as dense subgraph discovery. Table 3.1 presents the taxonomy of various event detection studies with their attributes. In the following, we discuss prior studies on event detection by segregating them based on the type and scale of events. First, we discuss large scale events which span over multiple documents, and subsequently, the detection of fine-grained events in text, and finally, we provide insights into event detection on multimodal data.

Event Detection at Multi-document Scale

Event detection techniques on news or social network streams can be put into two categories, namely *retrospective detection* and *online detection*. Retrospective detection deals with the detection of undiscovered events in the historical data. On the other hand, online detection performs the detection on the stream of stories in the online fashion. One of the earlier works in this domain, [Yang et al., 1999] present several approaches for detecting and tracking of events in News streams. They introduce two basic clustering approaches for event detection and tracking tasks in news streams. One of the approaches utilizes group-average-based hierarchical clustering, which maximizes the similarity among documents in a cluster by performing the merging in a bottom-up greedy fashion. The other proposed approach is an incremental clustering algorithm named single-pass clustering. It performs the clustering by considering a document at a time and assigns it to the most similar cluster already available if the similarity value exceeds a predefined *clustering threshold*; otherwise a new cluster is created with the document in consideration as the seed. In order to ensure the temporal coherence of clusters, the authors provide a linear decay function to introduce a time penalty in clustering.

Interestingly, [Li et al., 2005] introduce a probabilistic model for retrospective news event detection. They propose a multi-modal algorithm, which explicitly takes into account both the contents as well as the time information of news articles. In addition, an approach is proposed that allows to get the approximate number of events from articles count-time distribution. Here, an event is defined as a collection of persons, locations, keywords, and time. In order to model the content of news articles, a naive bayes classification model is used. Three NB models are used to model persons, locations, and keywords individually. Moreover, for timestamps, the authors use a Gaussian Mixture Model (GMM). Thus, the complete event detection model is the combination of three naive bayes unigram models and a GMM model.

In their work to explore the extraction of events in a stream of user-generated contents, [Ifrim et al., 2014] propose a technique based on the aggressive filtering and hierarchical tweet clustering. To this end, they employ two stage hierarchical clustering: once to cluster the tweets by the topic and secondly on the headlines resulting from the first step. A ranking mechanism is also utilized to rank and extract the trending events.

In [Angel et al., 2012], the authors attempt to discover real-time stories by identifying groups of tightly coupled entities (e.g., persons, products, locations, etc.). A dynamic graph is used to study the problem with entities as vertices and edges derived from co-

References	Approach	Event Types	Data	Remarks
[Feng et al., 2018]	LSTM + CNN	Fine-grained news events	News corpus ACE2005	English, Chinese, and Spanish languages
[Chaney et al., 2016]	Capsule model	Events in cables	National Archives' corpus	Employs topic modeling
[Gao et al., 2010]	Graph based clustering	Spam campaigns	Facebook	MD5 fingerprint, 2.08M posts with URL
[McClosky et al., 2011]	Dependency parsing	Biomedical	BioNLP'09	Reranking parser
[Li et al., 2013]	Structured Perceptron	Fine-grained news events	ACE 2005 corpus	Beam search for decoding
[Nguyen et al., 2016]	GRU	Fine-grained news events	ACE	Uses word embeddings
[Yang et al., 1999]	Hierarchical + incremental clustering	News events	News stories (Reuters, CNN)	Time penalty in clustering
[Li et al., 2005]	Naive Bayes + GMM	News events	News stories (CNN, MSNBC, BBC)	Events like earthquake, Halloween etc.
[Ifrim et al., 2014]	Hierarchical clustering	News events	Twitter	Entities for ranking clusters
[Chen and Roy, 2009]	Wavelet-based spatial analysis	Localized events	Flikr photos	Periodic and aperiodic events
[Xu et al., 2018]	Smooth nonnegative matrix factorization	Celebrity news events	People.com	Celebrity news and multimodal data
[Zhang et al., 2017]	Structured perceptron	News events	ACE 2005, ERE	Visual and textual features

Table 3.1: Taxonomy of various event detection techniques

occurrence observations of entities. The authors propose an algorithm for the maintenance of dense subgraphs under dynamic edge weight updates caused by streaming data. Also a few investigations are made to reveal impact of one story on another.

In [Matsubara et al., 2015], the authors study the dynamics of co-evolving activities/entities and conjecture that online activities compete for user resources (attention, money, etc.) like various species in ecosystem for food. In order to model the ecosystem of the Web, a non-linear dynamic model is presented to mine large-scale co-evolving online activities. Moreover, the authors present a parameter-free algorithm to fit the model.

A capsule model in order to detect and characterize events in the U.S. State Department diplomatic cables (confidential message exchanges) is presented in [Chaney et al., 2016]. The proposed model is primarily intended for human interpretability rather than forecasting. In this study, an event is modeled as the temporal deviation from the typical behavior or in other words, the interaction between entities that deviates from their usual interaction pattern. In order to detect an event, a measure is defined that quantifies the "eventness" of a time interval by utilizing the posterior expected values of event topics. The experiments are performed on the National Archives' corpus of over 2 million U.S. State Department cables from the 1970s[16].

Furthermore, [Gao et al., 2010] introduce a model to detect and characterize spam campaign events in online social networks. They employ a graph where nodes are posts by social network users and edges signify the similarity between posts. Two nodes are connected if they share the same URL (Uniform Resource Locator). Further, a MD5 hash algorithm based fingerprint is computed on the description of each of the posts. Any two posts are considered similar if there is a match of more than 19 bytes in their fingerprints. Now, the problem is reduced to finding the connected subgraphs. Subsequently, a graph based clustering algorithm generate clusters of spam campaign posts that correspond to a potential component of some spam campaign.

Fine-grained Events Detection in Text

The task of fine-grained events detection deals with extraction of events at a smaller level such as in a sentence. [Feng et al., 2018] propose a hybrid neural network architecture for language independent detection of events. Here, the authors combine the Long Short Term Memory (LSTM) units with the Convolutional Neural Network (CNN) in order to construct a neural network model that does not depend on manually hand-coded features but learns features that are most relevant to a particular language and the task (i.e., the event detection). The model learns to utilize the trigger words such as "release", "play", etc. to extract the events from sentences in different corpora of English, Chinese, and Spanish language news articles.

In [McClosky et al., 2011], the authors formulate the problem of event extraction as dependency parsing by considering the tree of event-argument relations. The motivation behind this is to capture the properties of both the nested and the flat event structures rather than working on individual events locally and independently. The authors utilized a logistic regression classifier with L2 regularization for detecting the event anchors. Once the event anchors and named entities are extracted from a sentence, a dependency

[16] http://history-lab.org/

parser (specifically, MSTParser[17]) is employed to generate the dependency links between them. Finally, a two-step re-ranking parser retrieves the dependency tree with highest score that incorporates the arbitrary global properties. In contrast to the pipelined approaches (such as aforementioned), [Li et al., 2013] tackle the problem in the holistic manner. They introduce a framework which jointly performs the structured prediction for triggers and arguments. In addition, the proposed model provides mechanisms for incorporating global features. [Nguyen et al., 2016] also proposed a joint event extraction model based on Recurrent Neural Networks (RNN). They employ word embeddings with a Gated Recurrent Units (GRU) model performing the event extraction task jointly, thus, improving over the model proposed by [Li et al., 2013]. [Liao and Grishman, 2010] propose the use of document level cross-event inference for improvement in the extraction of events. They train two MaxEnt (Maximum Entropy) classifiers, one for the classification of document-level event triggers, and the second for tagging the possible event arguments.

Events Detection in Multimodal Data

In the following, we discuss studies on event extraction in data with heterogeneous and multimodal characteristics. [Xu et al., 2018] conduct experiments on the click-through data for the detection of events. They aim at automatically detecting and creating chronologically ordered storyboard of events from query log data. The motivation behind it is that query logs directly capture the people's interests. For that purpose, they propose a framework to detect events by using the smooth non-negative matrix factorization technique, and retrieve relevant images corresponding to the events. A ranking function is introduced to highlight the social events, which exploits various information such as query semantics, temporal correlations and mappings from query logs. Finally, once the social events have been detected, commercial search engines are employed to retrieve images. Filtering/Selection of relevant images is done by using several local and global features as well as the image similarity (block-based intensity histograms).

In [Chen and Roy, 2009], a technique is presented that is based on wavelet-based spatial analysis in order to detect localized social events. The proposed approach works by first detecting the event-related tags using wavelet transform on their temporal and spatial distribution. It is followed by the event generation where the tags related to individual events are clustered. Finally, a set of photos is retrieved for each of the generated events represented by tag clusters. They conduct experiments on photos posted on Flickr[18] and report results on periodic as well as aperiodic events.

In an interesting study by [Rozenshtein et al., 2014], the authors generalize the event detection problem over activity networks (e.g. social network, sensor network), and define an event as subset of vertices in the graph that are close to each other and show an unusual high activity. Furthermore, the problem is formulated as MaxCut problem where vertices are given weights to represent activity level.

In [Zhang et al., 2017] an approach is introduced, which incorporates visual knowledge along with textual knowledge from the text documents in order to improve event extraction. The explicit visual information from images linked with a text document can help in

[17] https://www.seas.upenn.edu/ strctlrn/MSTParser/MSTParser.html
[18] https://www.flickr.com/

disambiguation of text-only modalities, and provide an augmentation of the feature space for the event extraction model. To this end, they introduce a structured perceptron model, which integrates visual and textual features. A visual repository serves as the source of explicit background knowledge and facilitates the disambiguation of the event type and its arguments to the correct instance. In [Li et al., 2016], the work explores a multimodal approach for the focused construction of knowledge base related to events. The authors utilize visual information regarding events to populate structured ontologies. A corpus of weakly supervised image-caption pairs is exploited to detect the visual components of events and automatically name them.

It should be noted that all the aforementioned techniques deal with only the detection of events at different level of granularity. They do not consider impact or analytics aspects of events. We work on events having societal significance and do not have localized fine-grained nature. In this thesis, our primary focus is not the detection of events but the prediction and analysis of their impact. So before discussed techniques are not direct competitors to our research work.

3.1.2 Event Impact Analytics

Related work to event impact analytics include mining Web data to analyze and predict the mass behavior in our society. In [Kallus, 2014], the authors investigate predictive signals in Web data collected from 300,000 sources in order to predict the upcoming societal events. Recently, social media have witnessed a tremendous interest for event analytics because of its real time nature. Wikipedia has been intensively studied in its nature as a collaboratively curated encyclopedic information platform [Ahn et al., 2011, Whiting et al., 2014, Freire et al., 2016, McIver and Brownstein, 2014, Osborne et al., 2012]. Various studies have been conducted on the Wikipedia data for detecting the trending topics/pages. Most of such studies consider the page views statistics for predicting trends. In [Osborne et al., 2012], the authors used Wikipedia page views data in order to aid event detection on Twitter. Whereas [Fetahu et al., 2015] investigate event and entity related information flow between news and articles on Wikipedia. furthermore, in [Whiting et al., 2014] the usability of Wikipedia as a source of temporal event information is explored. Although they do not conduct any extrinsic studies, they provide insights into various signals (such as page edits, page views, etc.) that can be exploited for disparate research problems. Also, Twitter has been widely used for detecting the societal events and popular trends [Petrović et al., 2010, Atefeh and Khreich, 2013]. [Petrović et al., 2010] present an algorithm based on locality-sensitive hashing to detect first story from tweets stream with low computational cost. An overview of some of the selected event impact studies is reported in Table 3.2. Various works on event impact analytics can be organized along the following categories.

Event Future Prediction

Recently, there have been growing interest in prediction of the future aspects of an event, for example, what other events or scenarios can follow, or where the event will become popular. In [Hashimoto et al., 2014], a supervised method for extracting the causal relationship between events is proposed that generates future scenarios based on semantic

References	Approach	Task	Event Types	Data
[Hashimoto et al., 2014]	Dependency parsing + SVM	Future scenario	Social events	Web crawl
[Muthiah et al., 2015]	Key phrase learning	Planned protest forecasting	Civil unrest events	Online news and social media
[Ahn et al., 2011]	K-means, LDA	Trends on Wikipedia	Societal events	Wikipedia pageviews statistics
[Radinsky and Horvitz, 2013]	Bayesian inference	Future events warning	Epidemic events	NYT news corpus
[McIver and Brownstein, 2014]	Poisson, LASSO model	Influenza-like illness prediction	Seasonal and pandemic diseases	Wikipedia page views
[Sakaki et al., 2010]	SVM, Kalman filters	Natural disaster event detection and location	Earth-quakes, Typhoons	Twitter
[Chakraborty et al., 2016]	Event based ARIMA	Socio-economic factor prediction	News events	Times of India archive
[Cadena et al., 2015]	LASSO	Civil unrest event prediction	Protests, Strikes, etc.	Twitter
[Piškorec et al., 2014]	Entity based document similarity	Relation b/w financial indicators and news	Financial market events	News from multiple sources
[Dos Santos et al., 2016]	Bayesian and spatio-logical inference	Event association analysis	Violent civil unrest	Twitter, GDELT
[Bastos et al., 2015]	Granger causality	Impact of ongoing events on social media, and vice-versa	Civil unrest events	Twitter, Facebook, and real world observation

Table 3.2: Taxonomy of various societal event impact studies.

relations, context and association features. For example given the causalities $A \rightarrow B, B \rightarrow C$, they generate an $A \rightarrow B \rightarrow C$ scenario by chaining the extracted causality. Here, candidates for the event causality are extracted by using dependency parsing (two event phrases in single sentence). Further, semantic relation based, context and association features are collected for each of the causality candidate. An SVM classifier is trained using these features, which learns to predict whether or not an event causality candidate is a legitimate causality or non-causality.

In [Radinsky and Horvitz, 2013], the authors present a model that can capture the evidential increase in the likelihood of occurrence of protests, disease epidemics and other societal events based on news stories and social media. A Bayesian probabilistic model learns probabilities of the future events conditioned on previous happenings. For generalization of the model, the authors use abstractions and leverage factual data from multiple resources using the liked open data platform.

The authors conduct a study to estimate the influenza-like illness activity in the United States with the help of Wikipedia usage in [McIver and Brownstein, 2014]. They introduce a Poisson model for estimation, which monitors the rate of usage of particular Wikipedia articles on a daily basis, and predicts the prevalence of influenza-like illness activity in real-time. In addition, another model is proposed, which is based on LASSO (Least Absolute Shrinkage and Selection Operator) regression analysis, and uses various predictor variables such as page view counts of Wikipedia articles which are selected with expert knowledge in the concerned domain.

In a real-time event detection study [Sakaki et al., 2010], social media are examined as a source for analyzing the societal happenings such as earthquakes, typhoons, etc. Here, the authors present an event detection and location estimation algorithm, which takes the user query as input, and provides the event occurrence and its location as output. An SVM classifier is built in order to classify social media (in this particular case Twitter) posts into a positive or a negative class representing whether a post is related to the event or not. Furthermore, in order to determine the accurate location and trajectory of the event, a particle filter model based on Kalman filters [Kalman, 1960] is employed. Interestingly, it has been observed that the natural disaster events lead to very little information diffusion in social networks as opposed to main-stream planned events, such as a "computer game launch".

In [Chakraborty et al., 2016], an event-driven system is introduced in order to predict the impact of news events on socio-economic indicators such as food prices. The authors argue that incorporation of event knowledge in prediction models, can provide underlying signals that derive socio-economic factors. The model is purely based on events extracted from news articles and does not use prior knowledge of subject. To this end, an event based ARIMA (Auto-Regressive Integrated Moving Average) model is introduced. The event based model outperforms the basic ARIMA model that does not incorporate any events related knowledge.

In a study on financial markets, [Piškorec et al., 2014] investigate the relation of cohesiveness in financial news stories published on the Web with respect to the market volatility. They propose the News Cohesiveness Index (NCI), which captures the collective behavior of financial news stories. The NCI captures the similarity within the news articles published on the Web. The news articles are represented

by individual vectors computed based on the entities contained. The authors report a strong correlation of financial indicators with the NCI when computed on financial news.

Crowd Behavior Analytics

The task of crowd behavior analytics deals with assessing the current trends of the society motivated by social events. In [Jatowt et al., 2015], a platform is presented to analyze and explore the past and future collective attention of Twitter users. This work examines the collective expectation and recall on social media with the help of various temporal mentions in natural language. The speed, scale and range of information spread on social network is investigated in [Yang and Counts, 2010].

In a study by [Bastos et al., 2015], the impact of social media on the ongoing real world social events (specifically, protests) is investigated and vice versa. This work assesses the Granger causality [Granger, 1969] between "on the ground" development of social events and the activity on social media. The authors perform a causal analysis of six different variables namely, tweets, posts, protestors, camped-out, arrested, and injured protestors. Some of these variables are observed in real world and others of social media (i.e., Facebook and Twitter, in this case).

Furthermore, in [Dos Santos et al., 2016], experiments on the association analysis of social events are performed in a big data setting. Three models based on distance-based Bayesian inference, spatial association index, and spatio-logical inference, are presented to perform the association analysis on sequences of events. The first model predicts the probability of re-occurrence of an event. The spatial association index quantifies the influence of one event on some other event. The spatio-logical inference based model explores whether or not an event can provide the explanation for other events.

The task of crowd mining is investigated in [Amsterdamer et al., 2013]. The authors, introduce a system named crowd miner that provides the facility of iteratively discovering the best questions to ask the crowd and extract patterns in received answers. The overall goal of the system is to find the association rules which are of high significance. In addition, this study presents the benchmark data set for systematic measurement of the crowd mining algorithms.

A prediction model that forecasts the social unrest events based on the prior activity cascades in social networks is introduced in [Cadena et al., 2015]. Here, the authors claim that "on the ground" occurrence of a social unrest event has a precursor activity cascade in social media (specifically Twitter here), which can be potentially detected. Activity cascades are detected via a graph derived from social media users and their posting activity. They employ a LASSO based logistic regression model, which takes the input features from activity cascades, and predicts probability of the future occurrence of a civil unrest event.

In [Muthiah et al., 2015], the forecasting of planned protests is modeled in the news and social media. The authors develop a system by using key phrase learning to extract mentions of the civil unrest events. They employ the probabilistic soft logic to reason about the location of extracted events. A time normalization technique is employed to resolve the future tense mentions regarding the planned protests.

In [Ahn et al., 2011], the authors propose a model WikiTopics that provides the textual explanation about "Which articles on Wikipedia are trending and for what reason?".

They make use of Wikipedia page views hourly statistics to select candidates for trending articles. Subsequently, a clustering algorithm groups these pages into coherent topics. Finally, the authors introduce several textualization techniques to provide the text explaining what does a particular cluster represents. The potential problem with these textualization techniques is that they are merely based on heuristics and can not always capture the notion of a topic in cases when there is high edit activity.

In the aforementioned studies, we have seen work dealing with prediction of an event future as well as trends in the society driven by events. However, spread and impact of an event from the inter-cultural perspective are not explored. Thus, we address the task of predicting event diffusion in foreign language communities. To the best of our knowledge no study has been performed on the entity-level event diffusion prediction.

3.1.3 Virality Prediction and Analysis

In this section, we survey a variety of studies on the virality analysis of Web contents. There are only a few related work addressing news virality in association with targeted countries. A study on the virality of tweets has been conducted by [Hansen et al., 2011] without considering the aspect of the named entities involved. They report that a news content with negative affect has more probability to become viral, and in contrast, the sentiment should be positive to achieve high popularity in case of a non-news content. They employ a Naive Bayes classification system for the sentiment analysis, and model tweet virality using a generalized linear model with assumption of binomial distribution.

STICS (Searching with Strings, Things, and Cats), on the contrary, provides a search engine that employs entity-level analytics to search documents, without providing country-specific analytics [Hoffart et al., 2014]. This work builds upon the knowledge base YAGO and uses the named entity disambiguation system AIDA to provide search functionality at the entity-level. [Jenders et al., 2013] introduce an approach in order to discover viral tweets, again without the notion of country-specific aspect. They investigate various factors such as number of followers, tweet length, hashtags, and sentiments behind the virality of tweets.

There are several studies, which consider the virality as a network phenomenon and, thus, model it as a social contagion. In one of these studies, [Weng et al., 2013] investigate the impact of structural trapping, social reinforcement, and homophily on the spread of social contagions in social networks. This work demonstrates that the future global popularity of an information content can be predicted by studying the community structure from its early spread.

In [Cheng et al., 2014], the authors investigate whether the future of information cascades can be predicted or not. They observe that temporal and structural features of a cascade trajectory are beneficial in predicting whether or not the cascade will keep on growing further. Apart from the temporal (speed of re-share, etc.) and structural (spread network properties, etc.) features, several content features are explored as well. However, any entity-level features have not been employed.

Furthermore, in [Fang and Ben-Miled, 2017], a study on virality of news articles also examines factors behind the virality, such as number of followers, tweet length, hashtags, etc. They frame the problem as a classification task, which has two output classes: slowly

fading news and rapidly fading news. For this purpose, they train a linear SVM classifier with unigram features from news content. It has been observed that news with negative emotions fades faster, in contrast to those news that carry positive sentiment, and terms about lifestyle and leisure.

[Keneshloo et al., 2016] perform experiments on predicting the popularity of news articles. They frame news virality prediction problem as a regression model where the output value represents likeliness of a news becoming viral. They utilize a variety of features such as article metadata, content, temporal, and social network features harvested within 30 minutes of a news article publication on the Web (Washington Post[19] is the news source here). [García, 2015] analyzes the geographical spread of news based on images in social media, and presents a visualization framework. However, this study also does not consider any features to capture the entity-level semantics.

In summary, we observed that the virality of Web contents with respect to different focused aspects (e.g., countries) is not much explored. Moreover, majority of the studies about the content virality do not harness semantic features derived from the entities contained. Thus, we exploit entity-level semantics and incorporate LOD to address the task of interlinking news articles and associated relevant countries.

3.2 Linked Open Data and Common Knowledge

In this section, we discuss studies that employ the structured and semi-structured knowledge via Linked Open Data (LOD) in improving over a variety of tasks. Also, we provide an overview of related work which leverage the semantics via entity-level analytics.

3.2.1 Knowledge Based Models

Using knowledge bases and other openly available semi-structured common knowledge like Wikipedia, has been proven to improve across a variety of tasks in information retrieval (IR) [Dalton et al., 2014] and natural language processing (NLP) [Strube and Ponzetto, 2006]. Recently, there have been tremendous efforts in building large-scale knowledge bases like DBpedia [Auer et al., 2007], YAGO [Suchanek et al., 2007, Hoffart et al., 2013], FreeBase [Bollacker et al., 2008], which provide structured knowledge along the spatial and the temporal dimension. Plethora of tasks from diverse disciplines have benefited by incorporating knowledge from LOD. Some of the recent studies are automatic selection of meaningful concepts for detection of complex events [Yan et al., 2015], disambiguation of named entities [Usbeck et al., 2014] or topic classification in the social media [Cano et al., 2013]. However, none of them addresses the issue of event diffusion on the Web.

In [Elberrichi et al., 2008], the authors utilize WordNet for the task of content classification by augmenting the feature set with concepts. In [Wang et al., 2016], the knowledge contained in ontologies is exploited by a heterogeneous information network to achieve better performance in text classification task. [Song et al., 2011] make the use of probabilistic knowledge bases to improve the task of short text clustering. They authors develop a Bayesian inference model to conceptualize the short text as well as words.

[19] https://www.washingtonpost.com

In [de Loupy et al., 1998], the authors employ WordNet for the query expansion and the classification of retrieved documents. Several query expansion strategies such as stemming and synonymy are investigated. They presents a clustering system which has both the hierarchical as well as the cluster based aspects, and intends to provide end users a structured map of retrieved documents. In [Huet et al., 2013], the authors utilize the knowledge base to mine the history via a large archive of news articles that spans over a huge span in temporal dimension.

However, ELEVATE framework makes extensive exploration of knowledge bases to discover the semantic connection of events to the foreign language communities (cf. Chapter 4). Also, our semantic fingerprinting approach (cf. Chapter 5) employs solely the type information of named entities derived from YAGO in order to create the semantic representation of a document, which can further be used for various tasks.

3.2.2 Entity-level Analytics

Entity-level analytics aims at leveraging semantic information via entities for improving the performance of higher level tasks in NLP, IR and Web science, etc. Here, we discuss several of these studies in the scope of our work. The incorporation of knowledge via canonical entities has shown to improve the task of event detection in automatic content extraction [Hong et al., 2011]. Here, the authors present a model that exploits entity types and background information for the detection of events and their types. In addition, the model provides better results by utilizing cross-entity inference, which is driven by the fact that entities of similar type will participate in similar type of events. In [Weikum et al., 2011], an extensive longitudinal analytics on large-scale Web archive data is performed. Here, raising the analysis to the entity level is a central component of investigation as it enables sophisticated exploitation of semantics across the time scale. Another beneficiary of the entity-level analytics is the computational fact checking. In [Ciampaglia et al., 2015], the authors present a method to automatically perform the fact checking by exploiting the knowledge graph. They argue that human-like fact checking can be approximated by finding the shortest path on a network representing concepts in claims. For this purpose, they utilize a knowledge graph of RDF triples derived from facts in DBpedia.

A model for sentiment mining that incorporates features about the named entities from an ontology, is presented in [Peñalver-Martinez et al., 2014]. As the task is related to opinion mining in movie reviews, the authors utilize a movie ontology that is used to impart entity-level features in the sentiment classification model. Moreover, another study shows that the task of future event prediction benefits from entity-level analytics at greater scale [Radinsky and Horvitz, 2013]. A variety of features are retrieved from multiple LOD platforms, and are combined to generalize the future event prediction model. The motivation behind this approach is to capture the semantics on an event, and perform the prediction derived from abstract understanding.

In conclusion, incorporating entity-level features, provides richer insights into contents, therefore, benefits a variety of aforementioned studies. In this thesis, we explore the importance of entity-level analytics and knowledge bases for a variety of tasks related to societal event analytics.

3.3 Type Classification Methods

Type classification has been studied on various levels of granularity. In the following subsections, we segregate and discuss prior research based on entity and document level of granularity, i.e., the entity type classification, and the document type classification.

3.3.1 Entity Type Classification

Assigning the most appropriate type(s) to individual entities is a crucial fundamental task, and is addressed by entity type classification. In [Fleischman and Hovy, 2002], the authors propose a supervised method to determine the fine-grained types of entities. They consider the local contextual information as well as the global knowledge derived from resources such as WordNet [Miller, 1995]. Using the combination of before mentioned features, they train multiple classification models such as k-NN, NB, SVM, and C4.5 decision tree. The decision tree model is reported to be performing the best.

FIGER (FIne-Grained Entity Recognition) is a fine-grained entity recognizer, which adapts the perceptron classifier to predict the fine-grained entity tags derived from Freebase [Ling and Weld, 2012]. This work employs a variety of features, which include contextual n-grams, PoS (Part of Speech) tags, syntactic dependency, and distributional similarity features. A CRF (Conditional Random Fields) based model is utilized for the segmentation, and for the assignment of types to the named entities a multi-class multi-label perception based classifier is adapted. They perform an extrinsic evaluation on the relation extraction task to measure the performance of their approach.

In [Rahman and Ng, 2010], a collective and hierarchical classification model is presented in order to determine the fine-grained semantic classes of nouns (total around 100). In this work, the authors make use of factor graphs to perform collective classification, and more than 30 features of disparate types such as morphological, semantic, grammatical, and gazetteers, etc. They introduce a collective classification method based on factor graphs, which enables to exploit the relational information.

HYENA (Hierarchical tYpe classification for Entity NAmes) is an entity type classification system on a very fine-grained type taxonomy [Yosef et al., 2012]. Here, the authors present a multi-label hierarchical classifier in combination with a meta classifier to predict very fine-grained types. The feature set includes features from context, grammatical structure, gazetteer, etc. It is worth noticing that HYENA does not use features from WordNet in contrast to the system proposed by [Rahman and Ng, 2010].

However, all of these approaches address type classification of individual entities in contrast to type classification of the overall document containing them. Hence, these approaches are related but not directly comparable to our work.

3.3.2 Document Type Classification

In one of the earlier works on document type classification, the use of various supervised machine learning models is explored in [Sebastiani, 2002]. The author examines various aspects of the document classification problem and survey the performance of several different machine learning based classification techniques.

References	Approach	Features	#Types	Data
[Fleischman and Hovy, 2002]	C4.5 decision tree	Word frequency, topic, WordNet	8	TREC9 database
[Ling and Weld, 2012]	Perception	Syntactic, Distribution similarity	112	Wikipedia corpus
[Rahman and Ng, 2010]	Factor graphs	Syntactic, WordNet, Gazetteers	92	BBN Entity type corpus
[Yosef et al., 2012]	Hierarchi-cal+Meta classifier	Context, Grammatical, Gazetteers	505	Wikipedia, others
[Joachims, 1998]	SVM	tf-idf	23	Reuters
[Elberrichi et al., 2008]	Naive Bayes	tf-idf	10, 20	Reuters, 20Newsgroups
[Johnson and Zhang, 2014]	CNN	Words with order details	2, 103	IMDB, RCV1
[Yang et al., 2016]	Hirarchical attention network	Word embeddings	5, 10, 10	Yelp, IMDB, Yahoo, others
[Joulin et al., 2016]	Hierarchical softmax + hashing trick	n-grams	5, 10, 10	Yelp, IMDB, Yahoo, others
[Allahyari et al., 2014]	Semantic associativity	Concepts	6	Wikipedia, Reuters
[Lilleberg et al., 2015]	SVM	tf-idf+word2vec	20	20Newsgroups
[Kim, 2014]	CNN	word2vec	2, 5, 2	MR, SST-1, SST-2, others
[Lai et al., 2015]	Recurrent CNN	Word embeddings	4, 20, 5	20Newsgroups, Fudan, SST, others
[Alec et al., 2016]	Ontological Machine Learning	Ontology	39, 12	Thomas Cook, DBpedia abstracts

Table 3.3: A comparative depiction of various type classification works on entity and document levels

Furthermore, in [Joachims, 1998], a support vector machine classifier is employed to learn with many relevant features. They represent a document as a tf-idf (term frequency - inverse document frequency) vector. A support vector machine classifier based on the RBF (Radial Basis Function) kernel is utilized to perform the categorization.

In [Elberrichi et al., 2008], the utilization of concepts from the WordNet ontology is investigated in combination with terms in documents to aid the classification task. A vector comprised of tf-idf features along with concept features from WordNet represents the document. They perform chi-square feature selection and employ a Naive Bayes classifier to categorize text documents.

In [Joulin et al., 2016], the authors present a fast and simple text classification technique based on a linear classification model with rank constraint. The authors utilize only the n-gram features to build the model. As there can be huge number of n-grams, a hashing trick is employed to access features faster. In this work, they also make use of hierarchical softmax to efficiently deal with the higher number of output classes.

An ontology-based text classification method [Allahyari et al., 2014] aims at classifying documents with respect to dynamically defined topics. The authors claim that their model does not require a predefined set of output classes and the corresponding annotated training set. The model can adapt to any given ontology of output labels, as it learns to map the thematic sub-graphs created from documents to the given ontology.

[Lilleberg et al., 2015] experiment combining the tf-idf features with word2vec [Mikolov et al., 2013] word embeddings. They present a classification model based on SVM that uses word2vec features weighted by tf-idf along with the original tf-idf features itself. The authors report that the model trained with this combination outperforms the models based upon individual set of features.

[Alec et al., 2016] use an ontology to associate documents describing an entity with their types. To do that, they automatically populate a domain ontology with respect to the information contained in each document. Then, they use an ontology-based machine learning tool to learn a definition for each type. If a document complies with the definition of a certain type, then it is classified as this type.

Recent approaches employ deep neural networks (DNNs) that perform better when given a vast amount of training data. Examples of such studies include, utilizing the word order of the textual data for document classification [Johnson and Zhang, 2014]. In this work, the authors adapt a CNN (convolutional neural network), which can operate on the input text of variable size. The proposed CNN model utilizes the word ordering information in the text to predict the document type.

A hierarchical attention network based on GRU (Gated Recurrent Unit) is presented by [Yang et al., 2016], which captures the hierarchical nature of documents. In this work, a two level attention mechanism is introduced, which operates at word and sentence level of a document. These two attention mechanisms focus on finding the important parts of documents to build a better representation that can be further utilized for classification.

In [Kim, 2014], a text classification model based on the convolutional neural network is proposed, which has been reported to provide encouraging results on a variety of classification tasks (such as sentiment analysis, ratings prediction, etc.). Here, the author makes use of word2vec pre-trained word embeddings, and presents several variants of static and dynamic word representation.

A classification system that employs both the RNN (Recurrent Neural Network) as well as the CNN is introduced in [Lai et al., 2015]. The proposed model has a convolutional module, which recurrently reads the sequence of words to obtain a document representation of higher quality. In addition, the authors introduce a max-pooling layer in the network that helps in determining which of the words in the text are of more importance to achieve the overall goal with higher accuracy.

In contrast to these work, our approach exploits entity-level semantics to build a concise document representation (i.e., the semantic fingerprint cf. Chapter 5), and does not have dependency on huge training data. In addition, we observed that there are not many document level type classification studies that target a large number of output classes. In this thesis, we explore the task of Web content classification with respect to a fine-grained hierarchy (with over 100 types).

Chapter 4

Event Diffusion in Foreign Language Communities

In this chapter, we explore the problem of event diffusion in foreign language communities. In order to achieve this, we introduce ELEVATE framework.

As previously mentioned in Chapter 1, with the availability of world-wide media coverage virtually all over the planet, happenings of various kinds have an almost global diffusion. This reaches from "relatively" minor incidents, e.g. an `Oscars 2017 mix-up`, with a mostly local perception up to globally registered events, such as the `Executive Order 13769` aka the "Muslim travel ban"[20]. The effect of happenings like the before mentioned can also be recognized as the "seismic waves" in cyberspace. While a smaller incident might trigger short term reactions in social media only, the impact of a more sweeping one eventually triggers a so-called "sh**storm" or might also find its way into Wikipedia.

[20] https://en.wikipedia.org/wiki/Executive_Order_13769

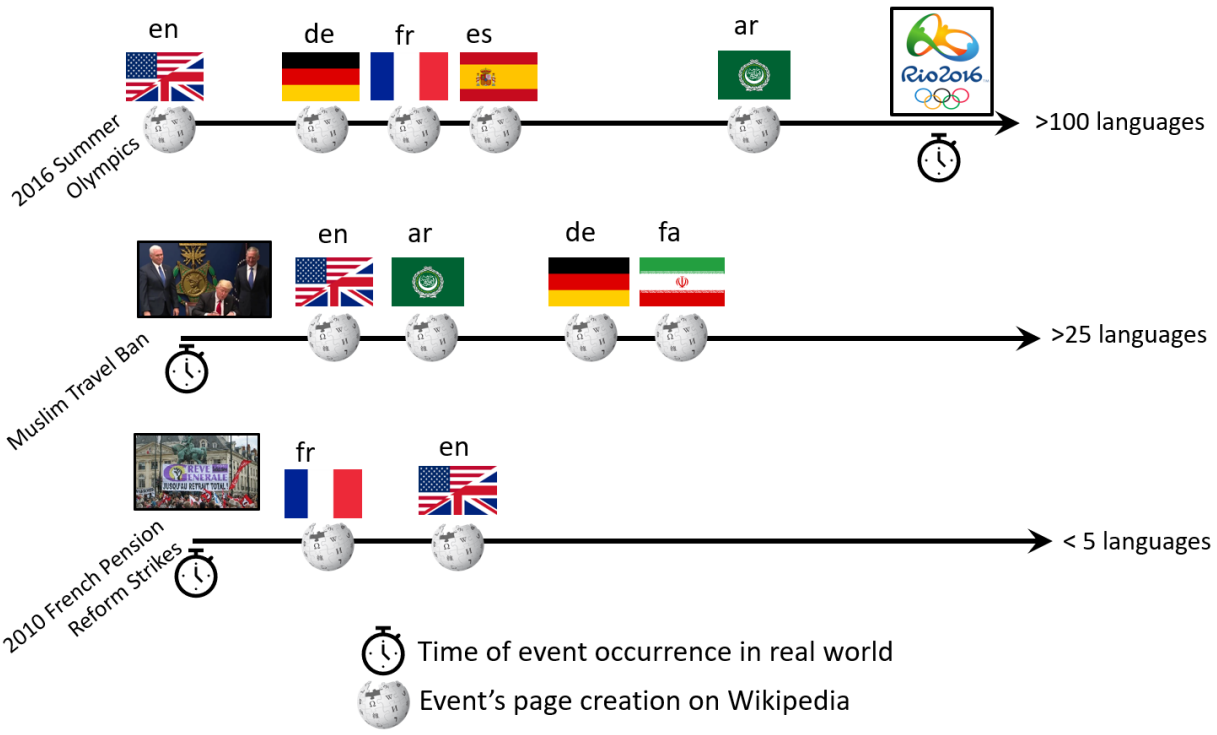


Figure 4.1: Diffusion of events to different language communities in Wikipedia; the clock symbol denotes the real world occurrence of an event

The aftermath of a societal relevant event is not necessarily limited to a single community, but might also influence others. For instance the before mentioned `Executive Order 13769` has so far been translated into more than 25 additional languages. Among these languages are - of course - also those, which are predominantly spoken in the affected countries, including Arabic. In contrast to small-scale events like `Oscars 2017 mix-up`, on the other end of the spectrum lies huge events like `2016 Summer Olympics`. Figure 4.1 depicts the diffusion of three different events into foreign language communities namely, `2016 Summer Olympics`, `Muslim Travel Ban`, and `2010 French Pension Reform Strikes`. It can be observed from the figure that depending on the nature, the original language and languages in future spread vary for one event to another. There can be huge events like `2016 Summer Olympics`, which spread to almost all major language of the world. On the other hand, there exist events like `2010 French Pension Reform Strikes` receive limited global attention. We consider an event as "societal relevant" if the user community of Wikipedia creates a corresponding entry. The key question, however, still remains: how to automatically assess the impact of societal relevant events with respect to their diffusion potential into foreign language communities? Our hypothesis therefore is: based on the origin of entities associated with an event, we are assuming a predictable information diffusion into the corresponding languages.

4.1 Conceptual Approach

With the availability of knowledge bases (KBs) such as DBpedia [Auer et al., 2007] or YAGO [Suchanek et al., 2007], and tools for interlinking the named entities in text documents with Linked Open Data (LOD) like AIDA [Hoffart et al., 2011] or DBpedia spotlight [Mendes et al., 2011], there have opened a multitude of opportunities to unravel the semantics of plain text document with the help of entities contained. A variety of knowledge can be incorporated from KBs via entity-level features. In our setting, the information utilized includes the country to which an entity is associated with and the predominant language of this country.

In this work, we introduce the ELEVATE (Entity-LEVel AnalyTics for Event diffusion prediction) framework. To this end, we investigate the diffusion of unscheduled, or, so-called "unforeseeable" events. These kind of events cover a wide spectrum, such as natural disasters, conflicts, political actions, etc. Specifically, ELEVATE semantically analyses Web contents by exploiting the information about the named entities contained. Figure 4.5 depicts the conceptual approach of the ELEVATE framework (cf. Section 4.3.2 for details). Based on the aggregated entity information it incorporates a diffusion model that has been trained and tested in order to accurately predict an event's impact onto foreign language communities. Other than existing methods, our approach is "purely" semantic, which derives all its knowledge from the YAGO knowledge base.

In summary, the salient contributions of the work presented in this chapter are:

- defining a prediction model of event impact,
- utilizing LOD for entity-level (semantic) analysis of the Web contents,
- experimentation with a multi-label classifier for identifying spread patterns in order to improve recall by simultaneously maintaining precision,
- a comprehensive experimental study on events in Wikipedia covering a time span of almost two decades demonstrating the high quality of our method.

4.2 Computational Model

A societal event $e \in E$ receives reactions on the Web in form of contents of varying nature like the Wikipedia pages, tweets, blogs and news stories etc. Moreover Web contents exist in multiple languages. We consider any publicly accessible resource on the Web with uniform resource identifier (URI) as a candidate Web content. The set of predominant languages on the Web is denoted by L. A Web content $w \in \mathcal{W}$ about an event e in language $l \in L$ is denoted by w_e^l. All the Web contents associated with an event e are represented by set W_e as shown in Equation 4.1. The named entities $n \in \mathcal{N}$ are the real world objects having abstract or physical existence e.g. persons, locations, organizations, etc. A named entity n in the language l is represented by n^l. N_j represents instances of the entity n_j in different languages as shown in Equation 4.2. The named entities

associated with a Web content w_e^l are language dependent and given by $\eta(w_e^l)$ as shown in Equation 4.3.

$$W_e = \{w_e^{l_1}, w_e^{l_2}, \dots w_e^{l_p} | l_1, l_2, \dots, l_p \in L\} \tag{4.1}$$

$$N_j = \{n_j^{l_1}, n_j^{l_2}, \dots n_j^{l_q} | l_1, l_2, \dots, l_q \in L\} \tag{4.2}$$

$$\eta(w_e^l) = \{N_1, N_2 \dots N_r | r \in [1, |\mathcal{N}|]\} \tag{4.3}$$

We define the graphs G_{link} and G_{sem} for our link-based and entity-level (semantic) prediction models respectively. The graph $G_{link} = (V_{link}, E_{link})$ is a directed graph with vertices comprised of the Web contents and named entities i.e. $V_{link} \subseteq \mathcal{W} \cup \mathcal{N}$. The graph G_{link} can be constructed using the inlinks and outlinks associated with the Web contents. It consists of two types of edges namely inter-language edges and mention edges. An edge is an inter-language edge if it connects two nodes representing the same named entity or Web content but in different languages. A mention edge connects a Web content to a named entity or vice versa, depending on what type of association between the Web content and named entity is considered. When the graph is being constructed using the outlinks then there will be a mention edge from the Web content to the named entity if the corresponding named entity is mentioned in this Web content. On the other hand, when the graph is being constructed based on inlinks then there will be a mention edge from the named entity to the Web content if the named entity is referring to it. So the edge set is $E_{link} \subseteq \{(\mathcal{W} \times \mathcal{N}) \cup (\mathcal{N} \times \mathcal{N}) \cup (\mathcal{W} \times \mathcal{W})\}$ in case of outlinks and $E_{link} \subseteq \{(\mathcal{N} \times \mathcal{W}) \cup (\mathcal{N} \times \mathcal{N}) \cup (\mathcal{W} \times \mathcal{W})\}$ in case of inlinks. Thus, edges connecting the two nodes representing the same named entity or Web content in different languages are bidirectional in nature.

The Graph $G_{sem} = (V_{sem}, E_{sem})$ is a directed graph with vertices comprised of the Web contents and named entities, and edges representing outlink-based association of the Web contents to named entities i.e. $V_{sem} \subseteq \mathcal{W} \cup \mathcal{N}$ and $E_{sem} \subseteq \mathcal{W} \times \mathcal{N}$. It is worth mentioning that the graph G_{sem} is less complex in structure than the graph G_{link} as it contains the outlink-based mention edges only and is bipartite in nature. The snapshots of the graphs G_{link} and G_{sem} at time $t \in T$ are denoted by G_{link}^t and G_{sem}^t respectively. Language $l_e^0 \in L$ denotes the Web content's language where an event e is being mentioned first. By $t_e^0 \in T$ we define the time of its premier appearance. For the sake of readability, in the following we refer to l_e^0 and t_e^0 by l^0 and t^0, respectively. For the same reason, subsequently G_{link}^t is referred by G_{link} as well as G_{sem}^t is referred by G_{sem}. The event spread $\psi(e)$ is the set of all languages in which there exist Web contents for an event e with $t^0 + \tau_\Delta$ specifying the time passed by after the event's first appearance (cf. Equation 4.4).

$$\psi(e) = \{l | \exists w_e^l \in W_e \text{ at } t^0 + \tau_\Delta, l \in L - \{l^0\}\} \tag{4.4}$$

Problem Definition. *Given a snapshot at time $t \in T$ of graph $G \in \{G_{link}, G_{sem}\}$ for an event e, predict the spread $\psi(e)$ i.e. the set of languages in which the event will appear eventually on the Web.*

We describe various approaches for event spread prediction in the following section. The generic prediction process involves four steps to compute the event spread $\psi(e)$. It starts with extraction of the available information for an event from a Web content W_e. This can be entities, inlinks and/or outlinks related to an event describing Web content

W_e. Then in the second step, additional information sources based on the model type are explored in order to discover an event's associations with different languages. In the third step, each language is scored based on the findings from previous step. In the last step, for each model the event spread $\psi(e)$ is computed by picking the most relevant languages from the scored candidates.

4.3 Event Spreading

Our hypothesis is, that the impact of societal relevant events can be assessed by the amount of reactions they trigger in the media as well as on the Web. Further, we assume that there are - like in the real world - inter-dependencies between the different actors involved and their impact on the resulting diffusion.

In order to predict the diffusion of event into foreign languages we investigate two conceptually different model types: link-based and entity-level (semantic). Each model exploits for a given event $e \in E$ the temporal snapshot $t \in T$ of graph $G \in \{G_{link}, G_{sem}\}$. Further, there is one language $l^0 \in L$, which can be identified as the premier language of event's appearance on the Web.

Link-based prediction: Based on the initial occurrence of the event e at time t^0 in language l^0, we can construct a link-based graph $G_{link} = (V_{link}, E_{link})$ at subsequent snapshots $t' = t^0 + \tau_{5\ days}$, $t'' = t^0 + \tau_{10\ days}$ and $t''' = t^0 + \tau_{20\ days}$. Thus, we obtain a representation of all other Web contents $w \in \mathcal{W}$ which is referring to or being referred by the event appearance $w_e^{l^0}$ in the premier language via a hyperlink. On top of the induced graphs at the time points t', t'' and t''' we gain per language information on the "importance" of an event based on the inlinks it receives and likewise in case of outlinks. Thereby, the languages with strong presence among the referring and referred Web pages $w \in \mathcal{W}$ for $w_e^{l^0}$ can be considered as candidates for the spread in near future. Figure 4.2 depicts the elements of the outlink-based construction of the graph G_{link} at a snapshot $t' > t^0$ in a graphical fashion. In a similar way, the graph G_{link} can be constructed using the inlinks as formalized in the previous section.

Entity-level prediction: As in the previous case, we construct graph based on the initial occurrence of the event e at time t^0 in language l^0 at the subsequent snapshots $t' = t^0 + \tau_{5\ days}$, $t'' = t^0 + \tau_{10\ days}$ and $t''' = t^0 + \tau_{20\ days}$. In contrast to before, we now construct a "semantically induced" graph $G_{sem} = (V_{sem}, E_{sem})$ which is solely induced via the named entities n mentioned in the event's Web page w. At this point, we are "lifting" the event to the entity-level only (without considering the Web graph structure). The core concepts of the entity-induced graph at a snapshot $t' > t^0$ are highlighted in Figure 4.3.

It is worth mentioning, that any implementation using the link-based prediction model requires "full world knowledge" in the sense that the entire Web graph must be known upon creation of the induced graph G_{link}. In contrast, the entity-level prediction model solely relies on the Web page w of an event e in the language of first occurrence l^0 and the named entities $n \in \mathcal{N}$ it contains. As such, the link-based prediction model has a clear competitive advantage over the entity-level model as it benefits from the inlinks and outlinks as a signal of "authority" and full contextual knowledge.

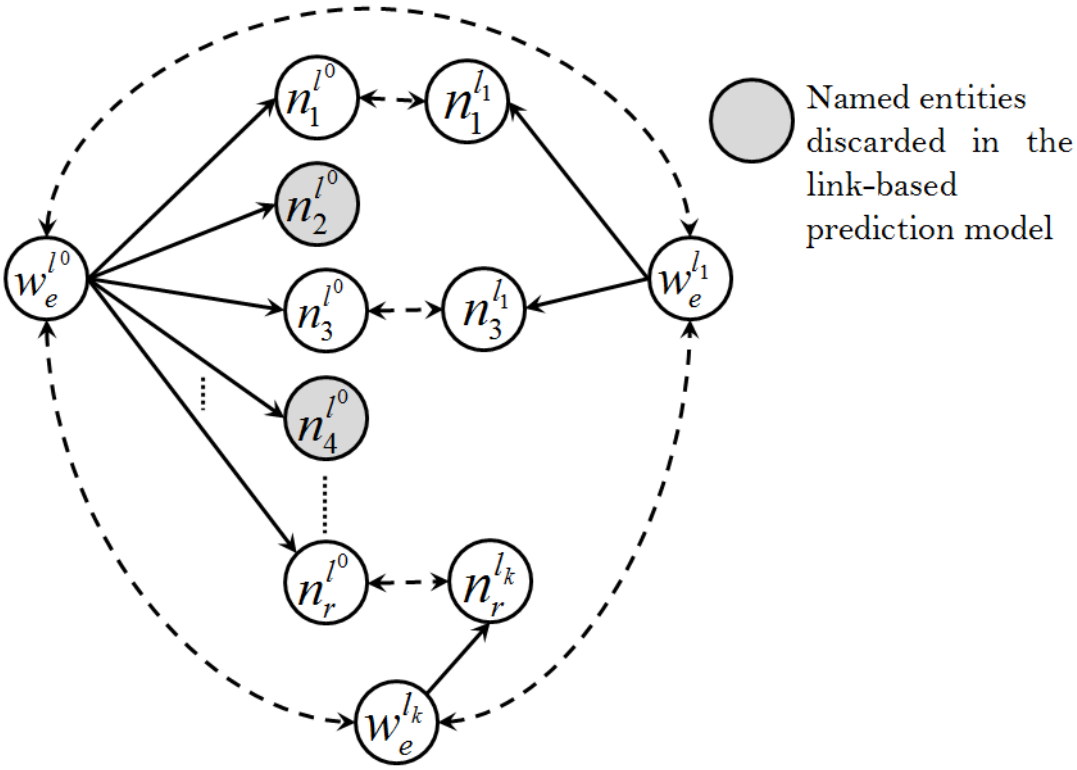


Figure 4.2: Outlink-based graph G_{link} at snapshot t'

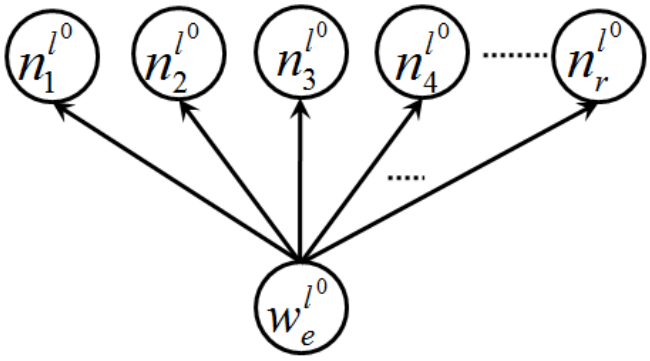


Figure 4.3: Entity-level (semantic) graph G_{sem} at snapshot t'

In the following we introduce in detail the aforementioned event spreading models. To this end, we present various different implementations in order to pinpoint specific characteristics of the underlying model.

4.3.1 Link-based Prediction Model

In the link-based prediction model we exploit the "neighborhood" of an event. To this end, we incorporate all pages connected to an event's Wikipedia page and induce a sub-graph associated with an event e_i. This subgraph is built upon the interwiki-language links to the same event in other languages ($w^l_{e_i}$) and the associated named entities ($\eta(w^l_{e_i})$). In order to express an event's importance, their are two notions of authority: outlink-based and inlink-based referral, which will both be investigated in the experimental section. The resulting sub-graph, thus, consists of a set of event related Web contents (in our case from Wikipedia) and entities linked to them.

The link-based prediction model now builds the aforementioned graph. In particular, we induce the connected sub-graph of an event e_i comprising all the corresponding event pages in different languages accessible via the interwiki-links, as well as the entities $\eta(w^l_{e_i})$ associated with all the before mentioned pages. In order to identify a coherent subsection of the cross-language connections and to eliminate outliers (e.g. pages that link to an event without any other connection to the events and/or entities of this event's neighborhood) we set the scope of the graph onto those entities n that are linked with at least two different language versions of the very same event e_i (formally $n \in \eta(w^{l_1}_{e_i})$ and $n \in \eta(w^{l_2}_{e_i})$ for $l_1, l_2 \in L$). Thus, all those named entities connected to just one language version of an event are discarded, because they do not contribute to the spread of event e_i. As a result, we obtain the graph G_{link} as highlighted in Figure 4.2 (bold lines indicate those neighborhood relations taken into consideration for prediction, while dashed lines indicate the corresponding translations of an event e_i into foreign languages incorporated for setting the scope). The grey nodes in Figure 4.2 are the named entities which are discarded because they are linked to only one language version of event e_i.

Based on the previous graph, we adapted several graph link prediction methods incorporating the node proximity measures. The underlying assumption is that the event (represented by its Wikipedia page), its translations into foreign languages and the entities (again connected via interwiki-links) form a "topological" cluster of this event. The intuition behind this approach is that an event e_i originating in the language l^0 - if found via an interwiki-link in a different language l_k - will cover a large amount of common context (in our context the named entities $n \in \mathcal{N}$). To this end, we identified and employed two similarity based measures namely common neighbors [Newman, 2001] and the Adamic/Adar [Adamic and Adar, 2003].

Common neighbors (referred to in the following as CN) can be formulated as in Equation 4.5 where x and y represent nodes in the graph, and Γ represents the neighborhood of a node (cf. Chapter 2). In our case, the nodes x and y are the Wikipedia pages corresponding to an event in different languages. The neighborhood of an event e_j first mentioned in l^0 is the set of named entities in its context i.e. $\eta(w^{l^0}_{e_j})$. Thereby, the common neighbors of an event e_j in the language of first appearance l^0 and the event's counterpart(s) in some other language $l_j \in L$ are the entities $n \in \eta(w^{l^0}_e)$ which have presence in

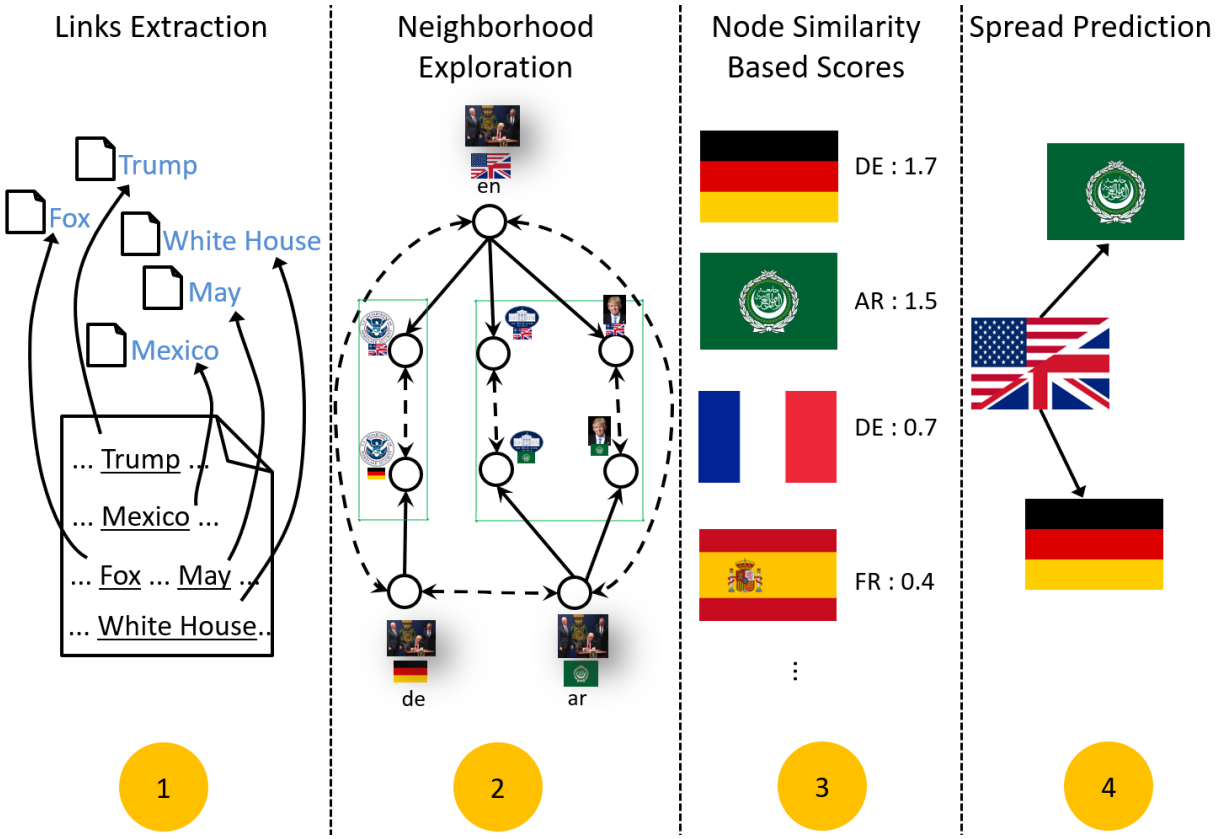


Figure 4.4: Conceptual approach behind the linked-based prediction methods

both the languages l^0 and l^j (formally $n \in \eta(w_{e_i}^{l^0})$ and $n \in \eta(w_{e_i}^{l_j})$ for $l_j \in L$).

$$CN(x, y) = |\Gamma(x) \cap \Gamma(y)| \tag{4.5}$$

Adamic/Adar (referred to in the following as AA) is formally defined by Equation 4.6. AA index is a weighted reformulation of the common neighbors similarity measure [Liben-Nowell and Kleinberg, 2007].

$$AA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log(|\Gamma(z)|)} \tag{4.6}$$

To this end, we score all the languages $l_j \in L - l^0$ for the event e_i using above described measures. It should be noted that existence of $w_{e_i}^{l_j}$ is not necessary for scoring a language l_j because all we need is the entities n from $\eta(w_{e_i}^{l^0})$ which exist in l_j. In addition, the aforementioned measures can be employed to compute the language scores based on the inlinks and outlinks separately.

Figure 4.4 depicts the conceptual approach behind the link-based methods for event spread prediction. As shown, there are four steps in the pipeline namely, links extraction, neighborhood exploration, node similarity based score computation, and finally the spread prediction. The first step, i.e., links extraction, deals with processing of an input document to extract the outlinks contained. In a subsequent step, the extracted outlinks are used

to explore the neighborhood of the Web content. In the third step, aforementioned node similarity based measures (i.e., CN and AA) are applied to compute scores for each of the possible languages in spread. The final step is spread prediction where the best candidates have to be selected from the all scored ones to generate the spread for the concerned event. We provide a detailed discussion about the spread prediction techniques that are employed in this work in Section 4.3.3. In a similar manner, the aforementioned pipeline can also work with inlinks instead of outlinks.

4.3.2 Entity-level (Semantic) Prediction Model

The semantic prediction model aims at exploiting information about the named entities mentioned in a Web content in order to derive their geographical provenance. The entity information is a valuable source in order to interlink Web contents with countries and, thus, languages. Figure 4.5 depicts the ELEVATE pipeline, which consists of the following four steps:

Step 1:
Initially, we build a sub-graph of G_{sem} corresponding to the Web content of event e in language l^0. To accomplish this, we process the Web contents in order to extract the named entities contained. To this end, we employ a state-of-the-art standard named entity disambiguation tool, AIDA [Hoffart et al., 2011, Yosef et al., 2011]. In the case of Wikipedia this process can be further simplified by exploiting the linked Wikipedia pages and directly mapping them on the corresponding canonicalized entity in YAGO [Suchanek et al., 2007, Hoffart et al., 2013].

Step 2:
Subsequently, we derive geographical information about the disambiguated entities associated with event e in previously built graph. For this purpose, we employ country and organization centric YAGO relations, such as `isLocatedIn` or `livesIn` (cf. Table 4.1 for the full list of relations). The process of extracting the countries associated with the entities is explained in detail in next subsection.

Step 3:
Based on the countries extracted in the previous step, we are now able to identify the languages associated with the entities mentioned and aggregate those scores. While there are in many cases several official languages associated with a country, we focus on the predominant language spoken in each of the countries mentioned. To this end, we automatically extracted from Wikipedia's "List of official languages by country and territory"[21] where for each country the first language is mentioned. The resulting mapping between countries and languages can be seen in Appendix A, and is also available on the project page of ELEVATE[22]

[21] https://en.wikipedia.org/wiki/List_of_official_languages_by_country_and_territory
[22] https://spaniol.users.greyc.fr/research/ELEVATE/

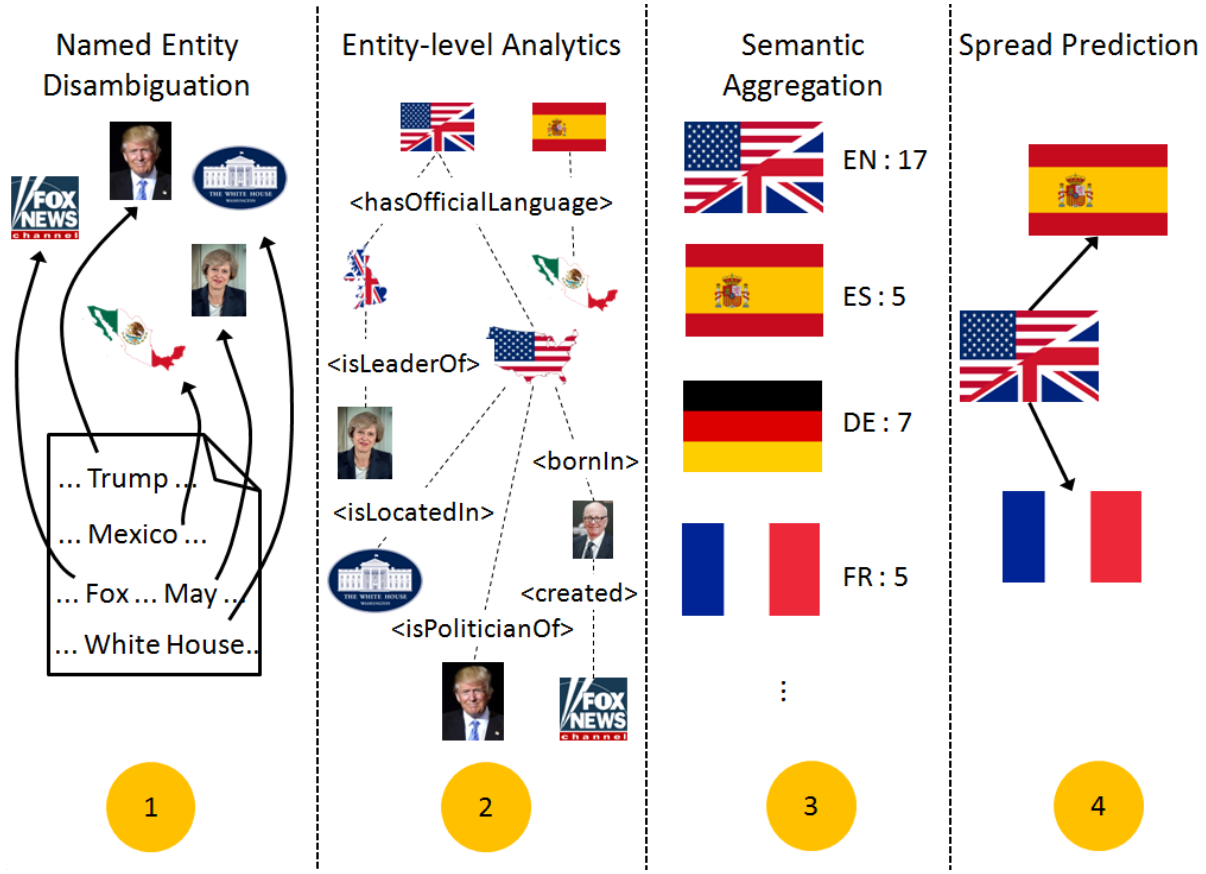


Figure 4.5: Conceptual approach of the ELEVATE pipeline illustrating the event on the "US-Mexico diplomatic crisis in 2017". (Graphical elements via Wikimedia Commons)

Step 4:
The final process then is the actual prediction based on the before aggregated entity-specific language scores. As a result, we obtain for each event e a language score vector. This vector is subsequently used in order to predict the "spread" $\psi(e)$ of this event into other foreign language communities. For that purpose, we tested and evaluated two different strategies: an adjusted thresholding based method and multi-label based classification approach. Both strategies are explained in detail in Section 4.3.3.

Country Extraction

Each language is scored based on its direct or implicit connection to the entities. The language extraction process comes into place, once the mentions of named entities have been mapped onto canonical entities. To this end, the connections of the entities to different languages are discovered using the semantic relations listed in Table 4.1 from the YAGO knowledge base.

The connections of entities to the languages can be direct by explicitly referring to the associated language or indirect by referring to the another entity from which country

Relation Type	YAGO Relation
Country centric	<isCitizenOf>
	<diedIn>
	<isLocatedIn>
	<isLeaderOf>
	<isPoliticianOf>
	<wasBornIn>
	<livesIn>
Organization centric	<owns>
	<created>
	<worksAt>

Table 4.1: YAGO relations used for entity-level analytics

and/or language related information can be found. To this end, we transitively traverse the before mentioned relations with a defined stopping criteria. Here, we apply two conceptually different strategies:

- Full exploration of languages
- Closest language discovery

When fully exploring the languages associated with the named entities mentioned, we traverse the entire knowledge base utilizing our semantic relations. This results in an exhaustive search and requires tracing of the explored path in order not to get stuck in a loop, i.e., similar to depth-first-search (DFS) with loop avoidance. The pseudo code in Algorithm 1 presents the implemented strategy. Naturally, this exploration strategy is susceptible to "semantic-drift" and might interlink several countries (and thus languages) that only have very loose ties with the original entity. In order to avoid this kind of unwanted expansion, we have also developed a strategy that aims at discovering the closest language, only. This approach is comparable to a breadth-first-search (BFS) with a stopping criterion once a country-related entity information has been found. The pseudo code in Algorithm 2 presents the implemented strategy. It is worth to mention that the algorithm does not immediately stop once a country-related entity information has been found, but further explores all other so far discovered entities "on the same level" for corresponding information. This has been done in order to treat entities on each level similarly and avoid misconception due to randomness in expansion. As a result, it is well possible that this algorithm also discovers more than one (different) language information per entity. Apart from above stated two schemes, a basic approach of directly mapping the country name in the event's page to its major language is also experimented.

4.3.3 Spread Prediction

The final process of the above described prediction models involve computation of the spread $\psi(e)$ for each event e from all the scored candidate languages. For this purpose,

Algorithm 1 Full exploration of languages

```
Data: Entity n
Result: Related language set langs and their scores
Stack S= empty
 S.push(n)
 while S is not empty do
    entity = S.pop()
     if entity is already visited then
        Do not explore
    else
        if entity is a country then
            official_lang = entity official language
             if official_lang is already in langs then
                Increment official_lang score in scores
            else
                Add official_lang to langs
                 Assign official_lang score to 1 in scores
            end
            Mark entity visited
        else
            Explore Yago for entity with following predicates: (isCitizenOf, diedIn, isLocatedIn, isLeaderOf, isPoliticianOf, wasBornIn, livesIn, owns, created, worksAt)
             Push explored entities to S
        end
    end
end
```

Algorithm 2 Closest language discovery

Data: Entity n
Result: Related language set *langs* and their *scores*
Queue Q = empty
Initialize cutoff level *cutoff* = ∞
$Q.enqueue(n)$
while *Q is not empty* **do**
  entity = *queue.dequeue*()
  if *entity is already visited or entity level* $\geq$ *cutoff* **then**
    Do not explore
  else
    if *entity is a country* **then**
      official_lang = *entity* official language
      if *official_lang is already in langs* **then**
        Increment *official_lang* score in *scores*
      else
        Add *official_lang* to *langs*
        Assign *official_lang* score to 1 in *scores*
      end
    else
      if *cutoff is -1* **then**
        Explore Yago for *entity* with following predicates: *(isCitizenOf, diedIn, isLocatedIn, isLeaderOf, isPoliticianOf, wasBornIn, livesIn, owns, created, worksAt)*
        Enqueue explored entities to Q
      else
        Do not explore
      end
    end
  end
end

we employ two techniques namely, adjusted threshold, and multi-label classification which will be discussed in details subsequently. The link-based as well as entity-based approaches use the same techniques in final step.

Adjusted Threshold

The thresholding is used to compute the spread $\psi(e)$ by only selecting the top Θ languages from the available set of scored candidates. Here, specifying a suitable threshold value is crucial as it has a strong effect on the event's spread. Not relying on the heuristic values, we compute the threshold based on the average spread onto other languages in the ground truth. In order to avoid over-fitting we perform the k-fold cross-validation where we learn the threshold using $k-1$ folds and predict the spread for events in the remaining fold.

Multi-label Classification

Intuitively the aforementioned thresholding scheme has the risk of picking the irrelevant languages when there is low variance or no difference at all in the scores of candidate languages. When there is sparsity in the available features (i.e. hyperlinks and entities in our case) it is likely to be the frequent situation that the prediction models assign more or less the same scores to all the candidate languages.

Thus, we model the spread $\psi(e)$ computation for an event e as a machine learning problem and specifically multi-label classification where the labels represent the languages of spread. The scores assigned to different languages by the previously described prediction models are utilized as the features. The spread of the event can be in any of the language $l \in P$ where $P = \{L - l^0\}$. Let $x \in \mathbb{R}^{|P|}$ represents the feature vector for some event e then our goal is to compute the vector $f(x) \in \mathbb{R}^{|P|}$ where $f_i(x)$ determines the membership of x in the target class $i \in P$. We use the one-against-all problem transformation method to transform the multi-label classification into a set of binary classification problem instances. We employ the support vector machine [Cortes and Vapnik, 1995] for classification purpose which is considered as a strong choice for large scale multi-label classification. One binary classifier for the individual class is trained to separate the members of that class from the members of all others classes. Given a training set of feature vectors $X \in \mathbb{R}^{|E| \times |P|}$ and corresponding set of class-label vectors $Y \in \{0,1\}^{|E| \times |P|}$, a decision function f is built for each class labels by considering the instances belonging to that class as the positive examples and all the others as negative. The decision function $f_i(x)$ is formalized in Equation 4.7 for a class label i and the test instance x on binary training set of instance-label pairs $(x_j, y_j), j = 1 \ldots |E|$ where $x_j \in \mathbb{R}^{|P|}$ and $y_j \in \{-1, 1\}$.

$$f_i(x) = sign\Big[\sum_{j=1}^{|E|} \alpha_j y_j K(x_j, x) + b_i\Big], \alpha_j \geq 0 \tag{4.7}$$

$$K(x_j, x) = \phi(x_j)^T \phi(x) \tag{4.8}$$

Where b_i is a bias parameter for the class i and $K(x_j, x)$ is a non-linear kernel function defined by Equation 4.8. $\phi(x)$ defines a feature mapping from the input space to the feature space. To this end, one classifier is trained for each language which learns the

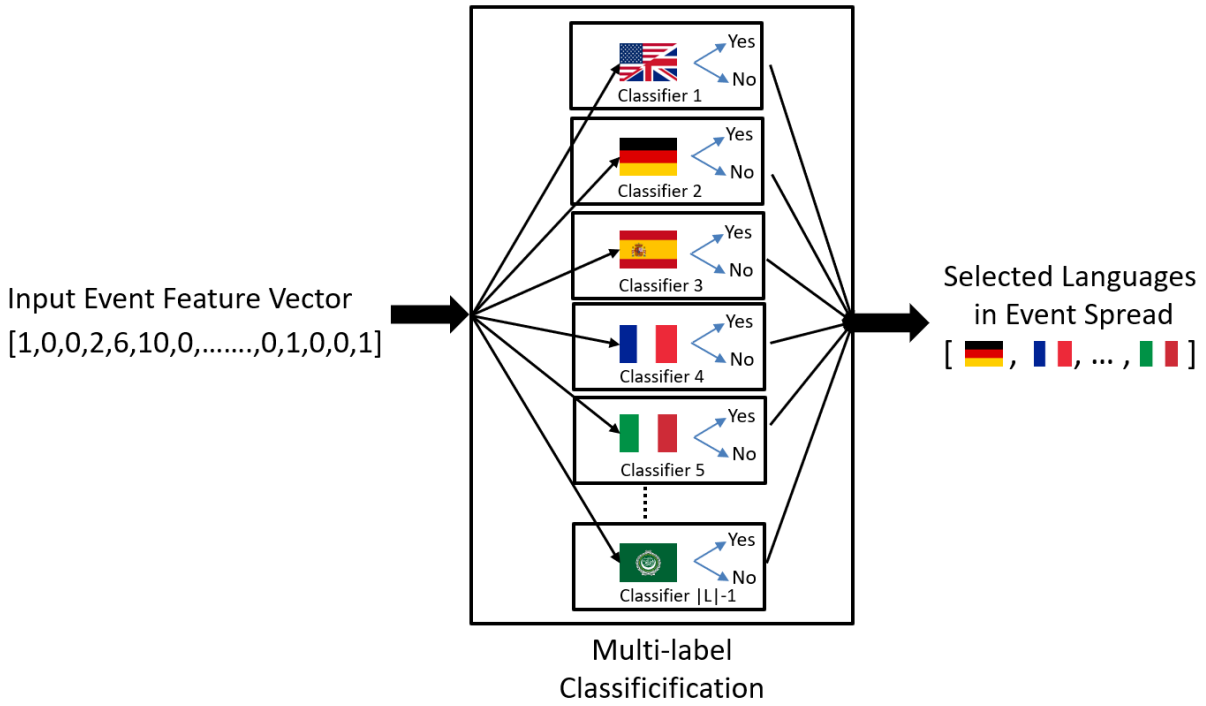


Figure 4.6: Multi-label classifier for event spread prediction using one-against-all problem transformation, i.e., a set of classifiers select the languages to be included in event spread

pattern of occurrence for this language based on the provided language score feature vectors. In total there will be $|L| - 1$ classifiers that will decide the spread of an event. Figure 4.6 depicts the process of multi-label classification for the event spread prediction. The input feature vector for an event is fed to each of the $|L| - 1$ classifiers corresponding to each of possible languages except the language of event origin. To this end, each of the classifiers take an independent decision considering whether the corresponding language should be in the event spread or not. Finally, decisions coming from different classifiers are merged to produce spread for the event.

4.4 Experimental Evaluation

We will now explain the setting of our experiments. To this end, we introduce the experimental set up before presenting the evaluation methods, experimental results and novel findings on the work.

4.4.1 Experimental Setup

In order to conduct our experiments we aim at investigating the societal relevant events. One, if not the, most important source for these kind of information is Wikipedia. To this end, we employ the Wikipedia portal of current events[23], which records events of

[23] `https://en.wikipedia.org/wiki/Portal:Current_events`

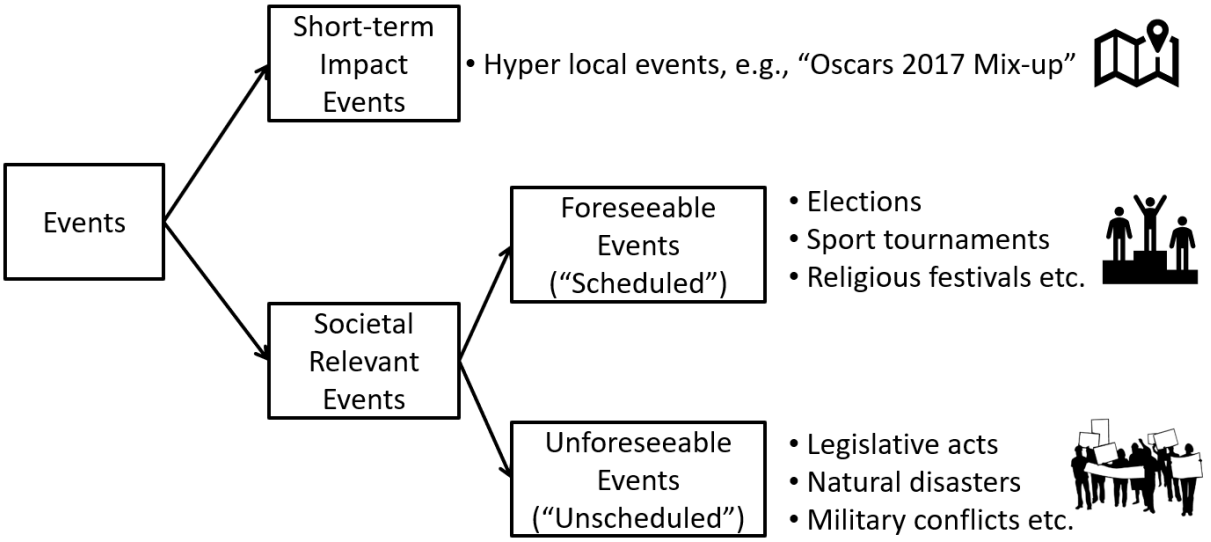


Figure 4.7: Taxonomy of different kinds of events from the perspective of societal relevance

the aforementioned kind around the world on a daily basis. From this Web source, we automatically extracted all current events linked to their own Web page. In the following subsections, we will in detail explain the creation of our experimental data set.

Event Types

Events can be categorized into two types from the societal relevance perspective namely, short-term impact events, and societal relevant events. Figure 4.7 depicts the taxonomy of events based on their societal relevance. The short-term impact events have a highly localized context and do not carry much significance in long term. The example of one such event can be "Oscars 2017 mix-up". On the other hand, there is a wide spectrum of the societal relevant events (also get covered in Wikipedia by individual pages). These events can be broadly classified into two different types:

- Foreseeable (predicted and scheduled) events
- Unforeseeable (unscheduled) events

The mentioned event types are conceptually very different. The foreseeable events include, but are not limited to, e.g. Olympic Games, periodic elections, Cosmic constellations, etc. These events allow a long look ahead period. As such, these kind of events are commonly anticipated and - as a consequence - listed in Wikipedia a long time before the actual event takes place. Hence, the event's details is subsequently refined over a long period of time (potentially spanning years). In contrast, the unforeseeable events cover a wide spectrum of events, such as natural disasters, conflicts, political actions, etc. These kind of events are, in general, non-predictable and hence get listed in Wikipedia temporally close to the actual occurrence. One can thus observe, that foreseeable are less dynamic than unforeseeable events. In our experiments we focus on those events that are "unforeseeable".

Unforeseeable Events Extraction

In order to extract unforeseeable societal relevant events, we employed Wikipedia's Current events portal[23]. This page lists the chronological order of events covered in Wikipedia. The contents relevant to our experiments cover each day in the time span form 2001 to 2016 retrieved via MediaWiki action APIs [24]. Spanning the 16 years of our experimental time frame, the Wikipedia "Current events Portal" contains links to more than 3,400 unique events. These events unroll across multiple languages which results in more than 43,800 Wikipedia event pages in different languages.

When using information from Wikipedia's "Current events Portal" one has to consider that the page not only contains the links to the events themselves, but also links to other related named entities (such as persons, countries, etc.). In order now to extract from all entities pointed to, those, that represent events, we employed Wikipedia's categorical system. The Wikipedia pages associated with the societal events tend to belong to the categories indicating time, geographical region and/or eventuality nature (e.g. *2017 in American politics*, *Conflicts in 2017* etc.). The events from year Y are highly likely to belong to one of the categories matching the regular expression "[[Y in %][% in Y]]". By employing this method, we are therefore able to identify the event pages (out of the set all entity pages linked from the portal page).

After having obtained the event pages based on the aforementioned process, the final step is the identification of those which are related to unforeseeable events. To this end, we consider an event as unforeseeable only if it is mentioned on current events portal in the same year as its page creation year on the Wikipedia. This filtering criteria captures the intuition that the event pages related to unforeseeable (unscheduled) events are created close to their actual occurrence in contrast to foreseeable (predicted and scheduled) events where pages are created far ahead of actual occurrence.

Experimental Data Set

From the previously described approach, we obtained more than 1,400 unforeseeable events from Wikipedia between January 2001 and May 2016. In order to perform our predictions, we consider those languages in Wikipedia that have a sufficiently large coverage. Therefore, we identified from the List of Wikipedias[25] those versions that contain more than 25,000 distinct pages (as of January 12, 2017). This lead to a total amount of 102 languages in our data set. Thus, predictions are possible for up to 101 languages (excluding the language of event origin).

It is safe to assume that the event pages in Wikipedia (in their nature as a Web encyclopedia) are monotonically increasing contents. This means, once an event in Wikipedia has been created, it is (in general) not going to be reversed anymore, but - potentially - refined and, thus, increasing in terms of contents, hyperlinks and entities contained. Table 4.2 confirms this hypothesis. Further, it is evident to see, that the interwiki-links among the event pages at the latest time point of our data set constitute "ground truth" of event-spread among languages. The latter observation also indicates that the knowledge

[24] `https://www.mediawiki.org/wiki/API:Main_page`
[25] https://en.wikipedia.org/wiki/List_of_Wikipedias

	5 days		10 days		20 days	
	Total	Average	Total	Average	Total	Average
Outlinks	76988	55	91134	65	105447	75
Inlinks	9386	7	13900	10	18384	13
Entities	23350	17	26635	19	30285	22
Countries	2433	2	2588	2	2945	2

Table 4.2: Temporal evolution of prediction parameters

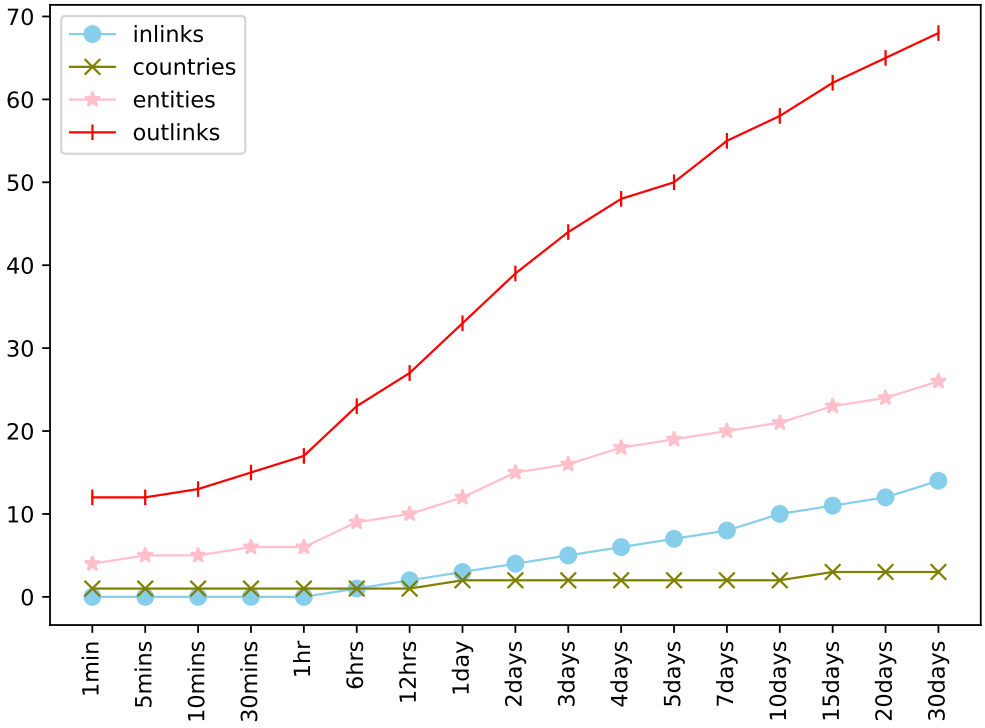


Figure 4.8: Graphical representation of the evolution of event parameters

about interwiki-links is a very strong indicator of language diffusion (which, in general, results in advantages of link-based approached over semantic approaches). Quality of the data set has been manually evaluated and is found to be over 95 percent accurate. The extracted ground truth events data set is available at the project page of ELEVATE[22].

4.4.2 Evaluation Methods

We performed extensive experiments on the dataset we introduced before. To this end, we conducted the 10-fold cross-validation experiments applying the previously presented (cf. Section 4.3) link-based as well as the entity-level (semantic) spreading models. We use the support vector machine library LIBSVM [Chang and Lin, 2011] for the classification purpose because of its open source availability. We evaluated different non-linear kernel functions namely the radial basis function and the polynomial kernel function on our

	5 days			
Method	Precision	Recall	F1	#PC
Random	0.07736	0.07691	0.07714	11288
$Inlinks_{CNC}$	0.23571	0.21023	0.22224	5045
$Inlinks_{Adamic}$	0.23553	0.21011	0.22210	5045
$Outlinks_{CNC}$	0.33301	0.42227	0.37236	10568
$Outlinks_{Adamic}$	0.33230	**0.42473**	**0.37287**	10568
$DirectMapping$	0.35574	0.11457	0.17332	1571
$Entities_{DFS}$	0.48845	0.22548	0.30854	4612
$Entities_{BFS}$	**0.53199**	0.19554	0.28597	2842

Table 4.3: Macro-average scores for the adjusted threshold based models after 5 days (#PC: number of predictions)

	10 days			
Method	Precision	Recall	F1	#PC
Random	0.07736	0.07691	0.07714	11288
$Inlinks_{CNC}$	0.25685	0.24077	0.24855	5875
$Inlinks_{Adamic}$	0.25685	0.24151	0.24894	5875
$Outlinks_{CNC}$	0.34840	0.45237	0.39363	10845
$Outlinks_{Adamic}$	0.34928	**0.45498**	**0.39518**	10845
$DirectMapping$	0.36171	0.12032	0.18058	1677
$Entities_{DFS}$	0.48580	0.23545	0.31718	4886
$Entities_{BFS}$	**0.53082**	0.20154	0.29215	3068

Table 4.4: Macro-average scores for the adjusted threshold based models after 10 days (#PC: number of predictions)

dataset, and choose the polynomial kernel because of its best performance among the two. To this end we utilize a polynomial kernel of degree 3 with γ and the cost parameter c tuned for the maximization of F1 score.

We distinguish in our experiments between the different time points of observation. In particular, we consider $t^0 + \tau_{5\ days}$, $t^0 + \tau_{10\ days}$ and $t^0 + \tau_{20\ days}$ after the first appearance of an event in Wikipedia. These time points have been chosen in order to capture the different dynamics an event attracts after its emergence: while events shortly after their emergence (e.g. $t^0 + \tau_{5\ days}$) have not necessarily received their full "attention", the coverage becomes fairly "mature" in due time (e.g. $t^0 + \tau_{20\ days}$). Table 4.2 depicts the evolution of various core parameters along the temporal dimension where average values are rounded off. It is worth to mention that - not surprisingly - for all of the observed parameters the absolute number of occurrences monotonically increases, which confirms our assumption about the archive-like character of Wikipedia.

	20 days			
Method	Precision	Recall	F1	#PC
Random	0.07736	0.07691	0.07714	11288
$Inlinks_{CNC}$	0.28541	0.27803	0.28167	6607
$Inlinks_{Adamic}$	0.28470	0.27854	0.28159	6607
$Outlinks_{CNC}$	0.36049	0.46988	0.40798	11049
$Outlinks_{Adamic}$	0.36182	**0.47766**	**0.41175**	11049
$DirectMapping$	0.36565	0.12337	0.18449	1770
$Entities_{DFS}$	0.48444	0.24076	0.32166	5087
$Entities_{BFS}$	**0.53368**	0.20823	0.29958	3234

Table 4.5: Macro-average scores for the adjusted threshold based models after 20 days (#PC: number of predictions)

	5 days			
Method	Precision	Recall	F1	#PC
Random	0.07736	0.07891	0.07813	11288
$Inlinks_{CNC}$	0.42002	0.19164	0.26320	5045
$Inlinks_{Adamic}$	0.41962	0.19146	0.26295	5045
$Outlinks_{CNC}$	0.34832	**0.33291**	**0.34044**	10568
$Outlinks_{Adamic}$	0.34756	0.33219	0.33970	10568
$DirectMapping$	**0.58498**	0.08311	0.14555	1571
$Entities_{DFS}$	0.36362	0.15172	0.21411	4612
$Entities_{BFS}$	0.49859	0.12819	0.20394	2842

Table 4.6: Micro-average scores for the adjusted threshold based models after 5 days (#PC: number of predictions)

4.4.3 Sensitivity Analysis

We perform a sensitivity analysis over the evolution of various parameters associated with events in the temporal dimension. Figure 4.8 depicts the growth of average number of outlinks, inlinks, entities, and countries for events in our ground truth. We measure the average counts at different time points from the creation of an event page. Specifically, we measure event parameters at eight different time points on the day of an event page creation, and then we increase the time interval of our observation as the event get older. To this end, we observed the parameters at 16 different time points i.e. after {1, 5, 10, 30} minute(s), {6, 12} hours, and {1, 2, 3, 4, 5, 7, 10, 15, 20, 30} day(s) from the time of an event page creation. We observed that outlinks receive the highest growth followed by entities, inlinks, and countries in respective order. We noticed that inlinks get a slow start as compared to countries, and countries lag behind inlinks as the time progresses. Also, it can be seen in the evolution graph from Figure 4.8 that entities stay higher than inlinks and countries count. This can be explained by the observation that entities mentioned in an event are usually more important to the event than the event to those entities. For example, the entity "France" is more important for some "Music Concert" event in France

	10 days			
Method	Precision	Recall	F1	#PC
Random	0.07736	0.07891	0.07813	11288
$Inlinks_{CNC}$	0.41260	0.21927	0.28636	5875
$Inlinks_{Adamic}$	0.41260	0.21927	0.28636	5875
$Outlinks_{CNC}$	0.35832	0.35152	0.35489	10845
$Outlinks_{Adamic}$	0.35924	**0.35242**	**0.35580**	10845
$DirectMapping$	**0.57662**	0.08746	0.15188	1677
$Entities_{DFS}$	0.35612	0.15744	0.21835	4886
$Entities_{BFS}$	0.48175	0.13371	0.20932	3068

Table 4.7: Micro-average scores for the adjusted threshold based models after 10 days (#PC: number of predictions)

	20 days			
Method	Precision	Recall	F1	#PC
Random	0.07736	0.07891	0.07813	11288
$Inlinks_{CNC}$	0.41305	0.24686	0.30903	6607
$Inlinks_{Adamic}$	0.41184	0.24613	0.30812	6607
$Outlinks_{CNC}$	0.36429	0.36409	0.36419	11049
$Outlinks_{Adamic}$	0.36564	**0.36545**	**0.36554**	11049
$DirectMapping$	**0.56836**	0.09098	0.15686	1770
$Entities_{DFS}$	0.35286	0.16266	0.22268	5087
$Entities_{BFS}$	0.47835	0.14000	0.21661	3234

Table 4.8: Micro-average scores for the adjusted threshold based models after 20 days (#PC: number of predictions)

when compared to vice versa. Also, there is naturally a delay because of the time required for the flow of information about the event. Thus, inlinks usually stay lower than entities in the initial period. Though, inlinks have the potential to overcome entities in longer term as an event can gain high influence with time.

4.4.4 Prediction Results

In this section we present the scores for several tried combination of the methods resulted from the underlying data set features employed and the aforementioned prediction models on top. To this end we have a total of four link-based models and three entity-level models. The evaluated link-based models are the common neighbors count with inlinks ($Inlinks_{CNC}$), Adamic/Adar with inlinks ($Inlinks_{Adamic}$), common neighbors count with outlinks ($Outlinks_{CNC}$) and Adamic/Adar with outlinks ($Outlinks_{Adamic}$). The entity-level models are the direct mapping of country names onto their major language, full exploration of all languages and discovery of the "closest" language(s) connected to the associated entities referred to as $DirectMapping$, $Entitites_{DFS}$ and $Entities_{BFS}$, respectively. Further the adjusted thresholding and multi-label classification are examined for

	5 days			
Method	Precision	Recall	F1	#PC
$Inlinks_{CNC}$	0.18634	0.20129	0.19353	7055
$Inlinks_{Adamic}$	0.21763	0.18534	0.20020	6028
$Outlinks_{CNC}$	0.27909	**0.38017**	0.32188	14169
$Outlinks_{Adamic}$	0.30896	0.37136	0.33730	13315
$DirectMapping$	0.31113	0.14553	0.19830	3538
$Entities_{DFS}$	**0.49333**	0.27747	**0.35517**	7179
$Entities_{BFS}$	0.45635	0.23128	0.30698	5413

Table 4.9: Macro-average scores for the machine learning approach after 5 days (#PC: number of predictions)

	10 days			
Method	Precision	Recall	F1	#PC
$Inlinks_{CNC}$	0.20312	0.22926	0.21540	8336
$Inlinks_{Adamic}$	0.22731	0.20412	0.21509	7011
$Outlinks_{CNC}$	0.28725	0.39046	0.33100	14320
$Outlinks_{Adamic}$	0.30479	**0.40219**	0.34678	14153
$DirectMapping$	0.31941	0.14587	0.20028	3529
$Entities_{DFS}$	**0.47678**	0.28490	**0.35667**	7509
$Entities_{BFS}$	0.46679	0.22763	0.30603	5408

Table 4.10: Macro-average scores for the machine learning approach after 10 days (#PC: number of predictions)

each method. We report the micro-averaged and macro-averaged precision, recall and F1 for three snapshot. Along with this, we report the number of classifications made by each approach denoted by $\#PC$ which is 11,057 in ground truth. The knowledge base (in our case YAGO) provides support for semantic relations among the named entities represented in 10 major languages only. There exist 1,049 events which first appeared in English and another 362 events in a non-English language. For the non-English languages we observe a resource scarcity in the knowledge base, which is due to YAGO's limitation on core facts extracted from 10 frequently used languages. Out of all events emerging in a non-English language, this results in 197 events in the languages not covered by YAGO. To overcome this shortcoming affecting the processing by semantic methods, we implement the spread onto English as default. In addition to the above methods, we also provide a "naive" baseline called *Random*. This method randomly selects a number of Θ languages (based on the average event spread) from the 101 languages under consideration. As *Random* does not use any temporally evolving event characteristics, the predictions made by this method are independent of time.

	20 days			
Method	Precision	Recall	F1	#PC
$Inlinks_{CNC}$	0.23301	0.25911	0.24537	9220
$Inlinks_{Adamic}$	0.25686	0.23574	0.24585	7687
$Outlinks_{CNC}$	0.28945	**0.41396**	0.34069	14533
$Outlinks_{Adamic}$	0.30112	0.40503	0.34543	14282
$DirectMapping$	0.32077	0.14912	0.20359	3755
$Entities_{DFS}$	**0.47549**	0.28420	**0.35576**	8210
$Entities_{BFS}$	0.46420	0.23649	0.31334	5702

Table 4.11: Macro-average scores for the machine learning approach after 20 days (#PC: number of predictions)

	5 days			
Method	Precision	Recall	F1	#PC
$Inlinks_{CNC}$	0.36173	0.22242	0.27546	7055
$Inlinks_{Adamic}$	0.41291	0.21693	0.28442	6028
$Outlinks_{CNC}$	0.29162	**0.36012**	0.32227	14169
$Outlinks_{Adamic}$	0.30079	0.34905	**0.32313**	13315
$DirectMapping$	**0.44686**	0.13779	0.21063	3538
$Entities_{DFS}$	0.32888	0.20584	0.25320	7179
$Entities_{BFS}$	0.37761	0.17819	0.24212	5413

Table 4.12: Micro-average scores for the machine learning approach after 5 days (#PC: number of predictions)

Threshold based Prediction

As observed in macro-average scores reported in Tables 4.3, 4.4, 4.5, the three entity-level approaches namely $DirectMapping$, $Entities_{DFS}$ and $Entities_{BFS}$ are best in precision across all three snapshots (i.e. $t^0 + \tau_{5\ days}$, $t^0 + \tau_{10\ days}$ and $t^0 + \tau_{20\ days}$). Moreover $Entities_{DFS}$ and $Entities_{BFS}$ perform better than inlink-based methods in F1, only losing to the link-based approaches $Outlinks_{Adamic}$ and $Outlinks_{CNC}$. Adamic/Adar performs slightly better than the common neighbors based approaches.

From the micro-average scores in Tables 4.6, 4.7, 4.8, it can be noted that $DirectMapping$ is best in precision due to the fact that it makes very concise predictions, because the presence of country names provides a strong clue for event spread in languages associated with this country. $Etities_{DFS}$ and $Entities_{BFS}$ perform lower than link-based methods in recall and F1 which is attributed to sparsity in entity information. It is worth noting that $Entities_{DFS}$ (full exploration of all languages) has best recall among the entity-level methods but loses in precision due to the noise induced from a "semantic drift". In contrast, $Entities_{BFS}$ makes the concise prediction by avoiding this drift and, thus, has better precision. In general outlink-based methods are better in F1, but use much more information that may not be available. As such, the link-based methods have a competitive advantage of having access to the full world knowledge in contrast to the

	10 days			
Method	Precision	Recall	F1	#PC
$Inlinks_{CNC}$	0.36972	0.26865	0.31119	8336
$Inlinks_{Adamic}$	0.39752	0.24294	0.30157	7011
$Outlinks_{CNC}$	0.29825	0.37230	0.33119	14320
$Outlinks_{Adamic}$	0.30460	**0.37578**	**0.33647**	14153
$DirectMapping$	**0.44092**	0.13561	0.20743	3529
$Entities_{DFS}$	0.32028	0.20970	0.25345	7509
$Entities_{BFS}$	0.36890	0.17392	0.23639	5408

Table 4.13: Micro-average scores for the machine learning approach after 10 days (#PC: number of predictions)

	20 days			
Method	Precision	Recall	F1	#PC
$Inlinks_{CNC}$	0.35944	0.28888	0.32032	9220
$Inlinks_{Adamic}$	0.40757	0.27310	0.32705	7687
$Outlinks_{CNC}$	0.30544	**0.38694**	**0.34140**	14533
$Outlinks_{Adamic}$	0.30647	0.38154	0.33991	14282
$DirectMapping$	**0.42104**	0.13779	0.20763	3755
$Entities_{DFS}$	0.30134	0.21607	0.25168	8210
$Entities_{BFS}$	0.36584	0.18193	0.24301	5702

Table 4.14: Micro-average scores for the machine learning approach after 20 days (#PC: number of predictions)

entity-level methods which are limited to the sparse entity information. The multi-label classification based event spread prediction is discussed in the following.

Multi-label Classification based Prediction

By transformation of the spread prediction problem as multi-label classification, the classifier learns to predict the languages in the event spread. As the features and output labels both are languages, the classifier is actually also learning the pattern of occurrence of a language with others. The entity-level methods has benefited more as compared to the link-based methods by this setting. Across different snapshots we observe an average improvement of 2%, 4%, 2% in macro-F1 and 6%, 3%, 3% in micro-F1 for $DirectMapping$, $Entities_{DFS}$, $Entites_{BFS}$ respectively when compared to the adjusted thresholding based performance. The driving factor behind this improvement is the improved recall as observed on Tables 4.9, 4.10, 4.11, and Tables 4.12, 4.13, 4.14.

The link-based methods do not improve except slight improvement in micro-F1 of the inlink-based methods. This behavior is generated because of the noise in features of the link-based methods. To explain it, let us take example of an event "AlphaGo versus Lee Sedol". This is an event which has a spread in around 10 languages. The feature set generated using the entity-level method $Entities_{BFS}$ has less than 10 languages

	Adjusted Thresholding			Multi-label classification		
	5 days	10 days	20 days	5 days	10 days	20 days
$Inlinks_{CNC}$	75	77	77	90	91	89
$Inlinks_{Adamic}$	82	84	82	93	93	92
$Outlinks_{CNC}$	79	82	84	94	96	96
$Outlinks_{Adamic}$	85	88	91	96	96	95
$DirectMapping$	59	59	59	87	88	88
$Entities_{DFS}$	57	57	57	93	93	92
$Entities_{BFS}$	58	58	58	90	89	91

Table 4.15: Number of predicted languages per method

which includes some of the most relevant languages to the event (Korean, Japanese, Chinese). On the other hand, the link-based method *outlinksAdamic* generates more than 70 features which includes irrelevant languages like Macedonian, Scottish etc. This makes the training data generated by the link-based methods more noisy and hinders the classifier to learn unambiguous patterns and leads to drop in the performance when compared to the adjusted thresholding approach.

4.4.5 Coverage of Languages in Predictions

In the previous subsections we highlighted the key results from our experiments. As indicated, we performed extensive results on a large number of events captured in English as well as non-English languages. From the threshold adjusted experiments we were able to confirm our hypothesis that entities are an excellent indicator in order to "locate" an events impact, which is reflected precision scores. In order to improve the recall without suffering a big loss in precision, we applied machine learning for multi-label based classification. In fact, this approach drastically increased the number of predictions by inferring common diffusion patterns (cf. Table 4.15 for details about the number of languages covered in the different prediction methods). Regarding the ground truth of 11,057 spreads, we consider the aforementioned approach as a suitable balance between precision and recall.

4.5 Findings on Event Diffusion

In this chapter, we have presented the ELEVATE framework in order to address the issue of predicting the spread of information into foreign language communities. We hypothesized that the spread of an event into foreign language communities depends on the named entities involved, and our study has shown it to be true. Our unique method introduces a novel approach of exploiting entity information from Web contents and harnessing the location related data for language related event diffusion prediction. Entity-level analytics with incorporation of semantics via LOD have been observed to be beneficial in gaining the understanding about the nature of an event. In conclusion, the most notable findings from the experiments with our ELEVATE framework are the following:

- predictions based on ELEVATE solely rely on semantic information without the need of exploiting neighborhood dependencies that require "full world knowledge",
- the ELEVATE framework scores best among all competitors on macro precision showing that entity information really capture the "semantics" of an event,
- when applying machine learning for multi-label classification, ELEVATE is able to significantly increase the number of predictions, which helps to increase recall by simultaneously preserving a high quality of precision.

In our comprehensive experiments on unforeseeable events covered in Wikipedia, we show that the exploitation of entity information allows a highly accurate prediction of spread into communities of different languages. In particular, ELEVATE doesn't require "full world knowledge" based on inlinks, which is required for the link-based approaches, but impractical in practice. By automatically learning typical patterns of diffusion we are furthermore able to increase the number of predictions (recall), by simultaneously preserving a high quality of precision.

As aforementioned, our approach is successfully able to reveal the spread of an event by deriving the semantic connections based on the named entities involved. However, there are cases when ELEVATE faces challenge in predicting certain language communities in the spread of an event. This arises from the fact that there exist certain users/editors on the Web who translate event pages to the language of their expertise as a hobby. Our system can not these perform prediction because of the non-existence of any observable semantic connection between the event and target language community. These beyond the scope scenarios are not observed very frequently in practice, thus, do not affect the overall viability of our system.

Chapter 5

Semantic Fingerprinting for Entity-level Content Classification

In this chapter, we pursue the task of web contents alignment onto a fine-grained type hierarchy. For this purpose, we introduce "Semantic Fingerprinting" approach.

"You shall know a word by the company it keeps!" is one of the most famous quotations by Firth [Firth, 1957]. Indeed, statistical linguistics backs Firth's theory that the context of a word gives a great insight about its meaning. Thus, in a nutshell, the local context of a word "defines the meaning". However, this leads to an extremely narrow contextualization. While this procedure might be helpful for type classification or even named entity disambiguation of the individual word (token), the semantic interpretation of the overall document remains unaffected. In particular, we intend to provide an efficient mean to allow a fine-grained content classification that goes beyond the classification of "basic" top-level types (such as, e.g. `person` or `event`). Hence, this work addresses the automatic classification of Web contents by "Semantic Fingerprinting". To this end, our goal is to align the Web contents with respect to a fine-grained type hierarchy as depicted in Figure 5.1. This task is inherently complex as it involves finding the most closely related type to the Web content from a taxonomy with large number of types and multitude of hierarchical relations between them.

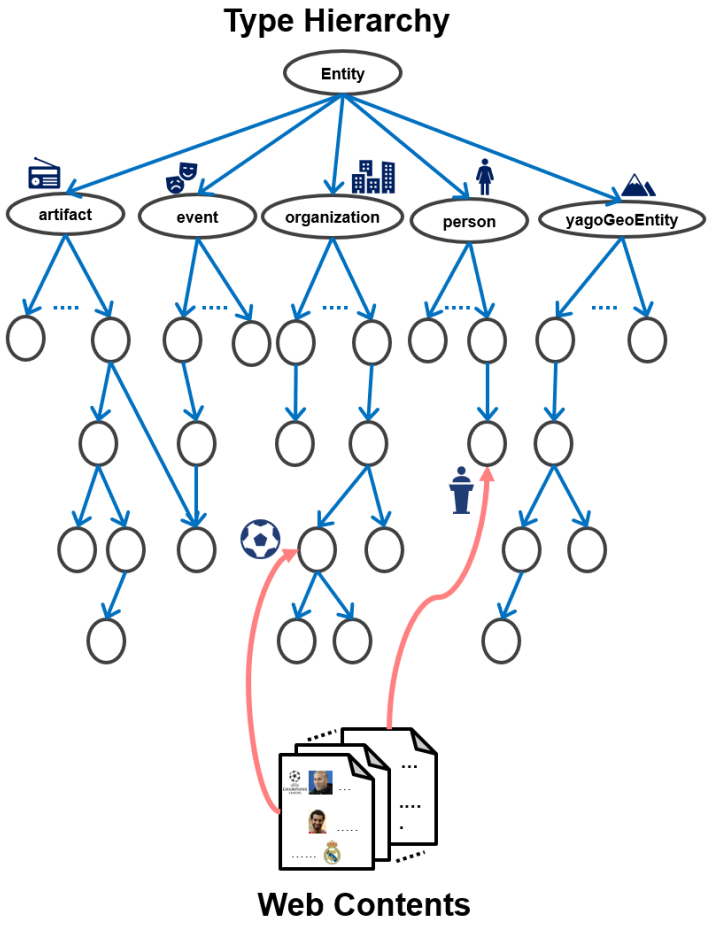


Figure 5.1: An illustration on aligning Web contents onto a fine-grained type hierarchy

In order to understand a Web content, a human requires to put the document in its context. To this end, the "crucial" information are identified, aggregated and interpreted. Key-point in this process is the identification and contextualization of the named entities contained that allow us to raise Web contents from "strings" to "things" [Hoffart et al., 2014]. Our research hypothesis therefore is: named entities contained in a Web content are "type-specific" and characteristic. In example, we postulate that Web contents to be classified, e.g., as type `(football-)club`, should not only be classified based on the "terms" contained, but more appropriately by the "things", e.g., `(football-)players` and `(football-)stadiums`. For that purpose, we harvest entity information and aggregate them as so-called "semantic fingerprints". These semantic fingerprints then allow us to efficiently classify Web contents.

In this work, we introduce "Semantic Fingerprinting" as a novel approach toward fine-grained entity-level content classification. To this end, we investigate classification of the Wikipedia articles based on their "inherent semantics" derived from the YAGO knowledge base [Suchanek et al., 2007, Hoffart et al., 2013]. This implies the exploitation and distillation of the entity-related information in Web contents for a subsequent classification according to a fine-grained classification scheme. In summary, the notable contributions of the work presented in this chapter are:

- a novel model of "Semantic Fingerprinting" Web contents,
- a machine-learning backed approach that allows a fine-grained classification of Web contents based on "semantic fingerprints",
- a comprehensive study on Web content classification based on the Wikipedia categorization scheme that shows the advantage of our approach in comparison to state-of-the-art competitors.

5.1 Computational Model

Apart from "google-style" indexing based on (freetext) keywords there has been recently a trend towards semantically classifying Web contents. With the emergence of large scale and fine-grained classification systems available via the LOD cloud there is an increased demand for interlinking (or simply classification) based on an underlying ontology, e.g., the Wikipedia category system.

Typically, a category prediction method involves collection of the document features followed by the application of a machine learning model such as logistic regression, Naive Bayes, etc. The most commonly utilized features for this purpose are bag-of-words. These representations are usually sparse in nature, and hardly captures the semantics of the document. We propose semantic fingerprinting, a method to represent the text documents based on the named entities it contains. This method exploits the type characteristics of the involved entities and produce a quality representation of the text document by capturing its semantic essence. We present our complete approach in the following subsections.

5.1.1 Type Hierarchies & Semantic Content Classification

The task of semantically classifying Web contents aims at assigning the proper type(s) /category(ies) associated. To this end, we utilize a fine-grained type hierarchy extracted from the YAGO knowledge base [Suchanek et al., 2007, Hoffart et al., 2013, Yosef et al., 2013]. This type-system is based on 5 top-level types (`person`, `location`, `organization`, `event` and `artifact`). In particular, we have chosen the 20 most populated (sub-)types per top-level type in order to automatically build our classification system based on "popularity". It is worth mentioning, that the resulting structure is a directed acyclic graph (DAG) and not a tree. It implies that certain (sub-)types might be associated with more than one super-type in order to express a context-depending facet of this type. This leads to a hierarchy consisting of 105 types. A graphical representation regarding details on the structure and types contained in the type hierarchy is available in Appendix C and on the project website[26]).

5.1.2 Type Score Vectors

In order to "semantically" summarize a document, we exploit the set E of entities contained, where each entity e_i belongs to a set of types denoted by $types(e_i)$. Let T be

[26] https://spaniol.users.greyc.fr/research/Semantic_Fingerprinting/105_hierarchy.pdf

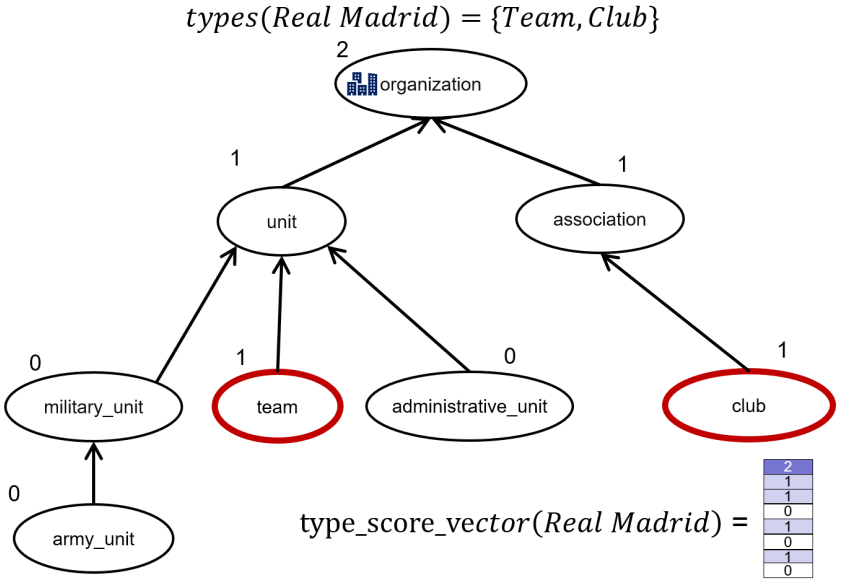


Figure 5.2: An example on a small fragment of our type hierarchy illustrating the computation of a type score vector for an entity

the set of all possible types, and the hierarchy H defining the relationship among them. For an entity e_i, the score of each of the types in T is aggregated by hierarchical upward propagation of the values of each type t in $types(e_i)$.

$$score(t) = score(t) + \sum_{k \in H(t)} score(k) \tag{5.1}$$

The children of a type t are given by $H(t)$. Using the recursive process given in Equation 5.1, a type score vector is computed for each entity e_i, which contains an aggregated score entry for each of the types in T. A semantic fingerprint $d \in D$ is the vector representation of the document, and is defined in the following subsection. Figure 5.2 illustrates the computation of the type score vector for the entity `Real Madrid` via a small part of our type hierarchy. The entity `Real Madrid` is directly associated with the types `team` and `club`, which are discovered using a knowledge base. To compute the type score vector, we assign a score of 1 to the nodes in our hierarchy representing types `team` and `club`. All other nodes are initialized with a score of 0. Now, we propagate the scores of children nodes to the parent nodes, and perform score aggregation as aforementioned. We keep on repeating the process until we reach the root node in the type taxonomy. The final scores are represented by a vector as depicted in Figure 5.2.

5.1.3 Semantic Fingerprint

As a semantic fingerprint, we define the vector d for a document, which is computed by summing the individual type score vectors for all the associated named entities $e_i \in E$ in the document/Web content:

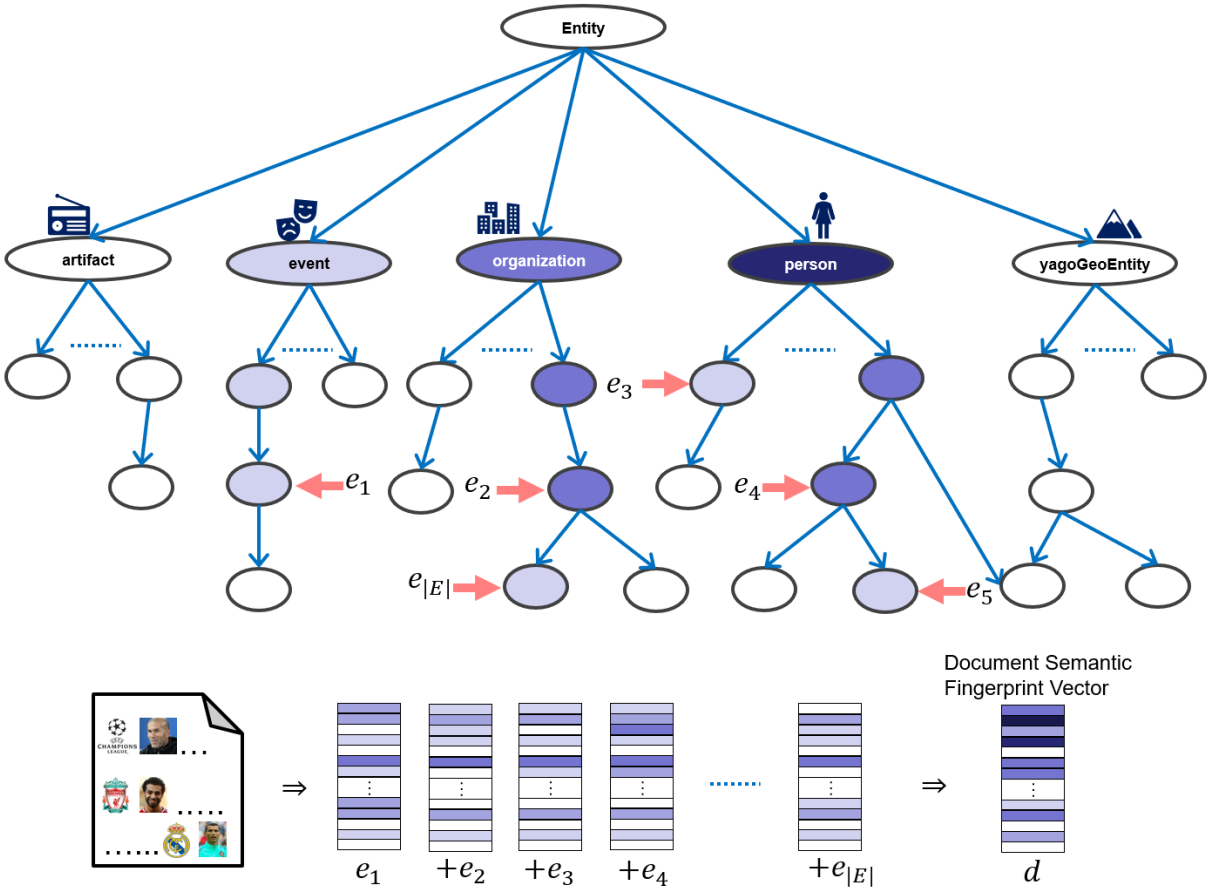


Figure 5.3: An illustration of the computation of semantic fingerprint for a document

$$\begin{bmatrix} d_1 \\ d_2 \\ . \\ . \\ d_{|T|} \end{bmatrix} = \begin{bmatrix} t_1 \\ t_2 \\ . \\ . \\ t_{|T|} \end{bmatrix}_{e_1} + \begin{bmatrix} t_1 \\ t_2 \\ . \\ . \\ t_{|T|} \end{bmatrix}_{e_2} + \cdots + \begin{bmatrix} t_1 \\ t_2 \\ . \\ . \\ t_{|T|} \end{bmatrix}_{e_{|E|}}$$

Figure 5.3 illustrates the generic process of computation of the semantic fingerprint for a document. As depicted, the process start with the computation of a type score vector for each of the individual named entities. The type score vector is of the same dimension as the semantic fingerprint, and is computed as discussed in previous subsection. The color intensity in the figure, is representing the magnitude of values. The document semantic fingerprint vector d is the summation of type score vectors. Figure 5.3 also depicts a sample of the type hierarchy with nodes colored with different intensities depending on their scores resulting from the aggregated type score vectors. It captures the intuition of how the sum over type score vectors aggregates the overall semantics of a document.

5.2 Classification via Semantic Fingerprinting

The originality of our approach is the use of semantic fingerprints that allows us to predict one or more fine-grained types to be associated with a given document. What we call a semantic fingerprint is a vector representing the entity types contained in a document. As such, we raise analytics to the entity-level, which has several advantages: firstly, it is compact as it consists of 105 types, only. As a result, the corresponding vector(s) are by orders of magnitude smaller than "bag-of-words" representations of the same content and, thus, more efficient to process. Finally, utilizing entity-level information means the exploitation of semantics. To this end, "Semantic Fingerprinting" is based on the following four consecutive steps. Figure 5.4 depicts the conceptual approach graphically.

1) Named Entity Recognition and Disambiguation
For each document to be classified, we extract and disambiguate the named entities $e_i \in E$. Conceptually, it is done, e.g., by AIDA [Hoffart et al., 2011] to disambiguate onto YAGO. In the case of Wikipedia, we obtain the entities directly via the mark-up. If the same named entity is present more than once in the content, its multiplicity does not count to the system as we consider only the unique entities.

2) Entity Type Hierarchy Computation
For each named entity e_i contained in a document, we derive the associated types from the underlying type hierarchy. In original, the named entities are labeled with as per type hierarchy of YAGO. As YAGO has huge type hierarchy, we map the type system used in YAGO to the type hierarchy considered here to filter out the infrequent types. By doing so, we are able to identify all associated (sub-)types out of the 105 before mentioned types (cf. Section 5.1.1).

3) Document Semantic Fingerprint
The semantic fingerprint of a document is represented by the vector consisting of an entry for each of the 105 types. Each value is the aggregated score for the entities of this type contained in the document. As discussed before the idea is to capture the core semantics behind the document and represent it by a vector of fixed length.

4) Type Prediction
Type prediction based on the semantic fingerprints requires the computation of a representative vector for each type $t \in T$. In order to do so, we construct for each of the 105 types a representative vector aggregated from 100 randomly drawn documents per type. Each document is processed as described in steps 1 – 3.

For the actual type prediction, we utilize a two-step process. This two step prediction process is needed to first identify the "characteristic pattern" of a type and then predict its "specific (sub-)type". First we use a classifier to learn and predict the top-level type. As mentioned in Section 5.1.1 the resulting type system is a DAG, so that (sub-)types with multi-parent nodes exist. Thus, a document can be typed by several top-level types. In this case, we randomly assign one of its top-level types in the training and the test sets. Once we have predicted the top-level type, we compute the cosine similarity as shown in

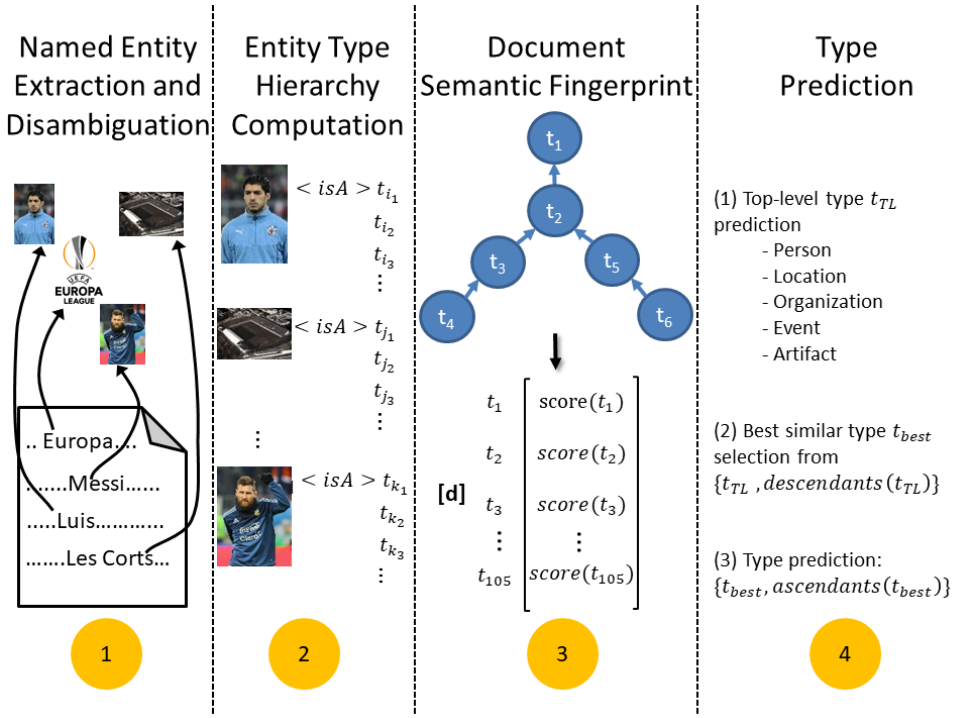


Figure 5.4: Conceptual approach by the example of a document of type `club`

Equation 5.2 between the representative vector of the document and the representative vector of each type that is a descendant of previously selected top-level type. We then select the one with the highest score and all its ascendant types.

$$cosine(X, Y) = \frac{\sum_{i=1}^{|T|} X_i * Y_i}{\sqrt{\sum_{i=1}^{|T|} X_i^2} \sqrt{\sum_{i=1}^{|T|} Y_i^2}} \tag{5.2}$$

5.3 Experimental Evaluation

In this section, we evaluate our "Semantic Fingerprinting" (indicated by *SemanticFingerprint*) method against two baseline methods "Naïve Bayes" [Manning et al., 2008] and "Elberrichi" [Elberrichi et al., 2008] (referred by *NaiveBayes* respectively *Elberrichi*). While "Naïve Bayes" is a common baseline used in document classification, the method of "Elberrichi" is a "semantic competitor" employing WordNet [Miller, 1995] on the most relevant noun phrases.

5.3.1 Evaluation Data Set

We use Wikipedia as the source of ground truth. In order to do so, we have chosen 10.500 random examples for training and 1.050 for testing. As Naïve Bayes benefits from prior probabilities, we maintained the ratio in the number of training examples accordingly. For the other two approaches, the training set consists of 100 random examples for each of the

105 types. The test set comprises 10 random examples for each type. Markup and stop words are removed during the preprocessing. In our *SemanticFingerprint* method, a random forest classifier is employed to select a single top-level type. It is trained with the full training set in each bag and computing the attribute importance with mean impurity decrease. This classifier achieves 68% accuracy on the top-level types.

5.3.2 Evaluation Strategy

We evaluate the performance of our approach with respect to Precision, Recall and F1. Figure 5.5, highlights various examples for type 'C' prediction. Example 1 is actually more specific (it is an example of 'D'). Example 2 is the same as 1 but it is also multi-labeled: it is typed as both 'D' (and its ascendants) and 'E' (and its ascendants). In both cases, we consider a prediction leading to the type 'F' (and its ascendants). In our data set, most cases are like Example 1, i.e., only a few of them ($\sim$ 10%) are multi-labeled. Thus, we do not apply multi-label prediction. In the following, we explain two different evaluation concepts that enable the assessment of various aspects of hierarchical multi-labeled data.

Full Evaluation
We consider all types mentioned in the ground truth and all prediction types. For Example 1, 'A' and 'B' are true positives, whereas 'A','B' and 'E' are true positives for Example 2 (cf. Figure 5.5).

Focused Evaluation
In this setting we predict an exact type. To this end, we consider as ground truth only this specific type and its ascendants ('C' and the nodes with black label in Figure 5.5) as ground truth. Thus, we assess all predictions against all types up to the level of the assessed node: level 3 for node 'C' in our examples, so we consider nodes on that level or above from the predictions. Moreover, in case of ground truth with multi-labeled types (such as Example 2), we remove from the predictions the correctly predicted types that are not the focus of the ground truth. In particular, in Example 2 'E' (level 3) is removed from the predictions because it is not considered as ground truth anymore, since the assessment is focused on 'C' and its ascendants.

5.3.3 Results and Discussion

We report the micro and macro averaged scores on the test set (cf. Chapter 2 for averaged scores). The results of macro-averaged as well as micro-averaged evaluation is summarized in Table 5.1 respectively Table 5.2. Our method (*SemanticFingerprint*) outperforms the other approaches in all performance measures and evaluation strategies. *NaiveBayes* performs the weakest, due to the complexity of the prediction problem and the lack of clearly discriminating features on the word-level. The second "semantic" method incorporating WordNet introduced by *Elberrichi* performs slightly better than *NaiveBayes*, but is still more than 5% weaker than our approach. In this setting, *Elberrichi* uses 200 features to represent a type as it is claimed to be best configuration by the authors. Thus, *Elberrichi*

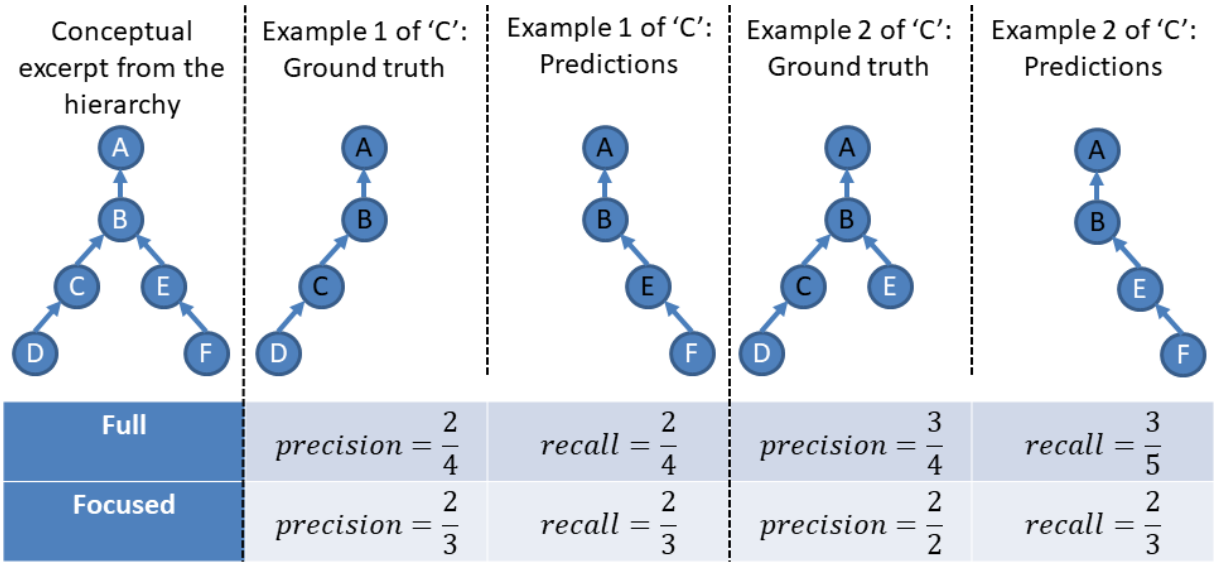


Figure 5.5: Full and focused evaluation

requires almost double the number of features as compared to *SematicFingerprint*. In addition, it employs both the terms as well as the concept frequencies in feature set and requires a computationally expensive feature selection step. *SematicFingerprint* uses feature vectors of size 105 to represent types and documents. Nevertheless, it gains over *Elberrichi* around 3.8% and 5.7% in macro-F1 for the full evaluation and the focused evaluation as well as around 2.6% and 2.7% in micro-F1. In summary, *SematicFingerprint* benefits from the compact and concise representation incorporating information derived from entity-level. All the training and test data sets used for different approaches in our experiments can be found in the associated zip-file[27].

	Full Evaluation			Focused Evaluation		
Method	Precision	Recall	F1	Precision	Recall	F1
NaiveBayes	0.52032	0.12026	0.19537	0.51651	0.19588	0.28404
Elberrichi	0.52421	0.42322	0.46833	0.53853	0.48325	0.50940
SemanticFingerprint	**0.59992**	**0.43766**	**0.50610**	**0.62063**	**0.52097**	**0.56645**

Table 5.1: Macro-average scores for document type classification

	Full Evaluation			Focused Evaluation		
Method	Precision	Recall	F1	Precision	Recall	F1
NaiveBayes	0.55185	0.10270	0.17317	0.54626	0.15866	0.24590
Elberrichi	0.56854	0.41848	0.48210	0.58719	0.50207	0.54130
SemanticFingerprint	**0.62545**	**0.42845**	**0.50854**	**0.63756**	**0.51318**	**0.56865**

Table 5.2: Micro-average scores for document type classification

[27] https://spaniol.users.greyc.fr/research/Semantic_Fingerprinting/data.zip

5.4 Findings on Entity-level Content Classification

In this chapter, we have introduced "Semantic Fingerprinting" as a novel method for entity-level content classification. Our study on Web content classification has proven that a document can be characterized by the named entities it contains, as hypothesized initially. By raising contents to the entity-level, we are able to capture the semantics concisely and efficiently. Based on extensive experiments on Web contents classified in Wikipedia, we have shown the viability of our approach and its performance gain against state-of-the-art competitors. The type information related to the named entities contained in a Web content have proven to be key in determining category of the Web content. It is observed that "semantic fingerprints" when assisted by machine learning can provide an effective solution for the content classification problem. The real life applications of our approach include automatically categorizing media articles, Web pages, books, etc.

However, there are still possibilities of further enhancements such as reducing misclassification among types that are inherently hard to separate. For example, the types `football player` and `club` are not clearly distinguishable when seen from the contained named entities perspective. Also, our approach might extend to classify arbitrary Web contents and provide mechanism to adapt to application-specific classification systems. Semantic fingerprinting has shown its efficacy and is not much affected by the before mentioned limitation as they are applicable to only a small number of special cases.

Chapter 6

Online News Virality Analytics

In this chapter, we address the problem of virality prediction for online news. We present ELEVATE-live framework for news virality prediction and visualization, which is an extension of the ELEVATE.

"Viral News" is a standing term that describes news that receives perception beyond average and, thus, spread at high speed and/or extremely wide. In particular, the Web allows a potentially global diffusion in almost zero time. However, different notions of virality exist in terms of speed, outreach, etc. with respect to the "importance" of a news article. While "importance" is still highly subjective and context - respectively - community dependent, the actors (named entities) involved are valuable indicators of the content's "inherent semantics". For instance, a report about the BREXIT and its consequences for Britain and its European partners is likely to contain named entities, such as politicians like `Theresa May` or `Emmanuel Macron`, organizations such as the `European Banking Authority` and the `European Commission` as well as cities respectively countries such as `Frankfurt`and `Luxemburg`. While country or city names are straightforward indicators for the importance of an article with respect to the mentioned place, it requires deeper semantics to derive this information for persons or institutions. With the emergence of automatically constructed large scale knowledge bases (KBs) such as

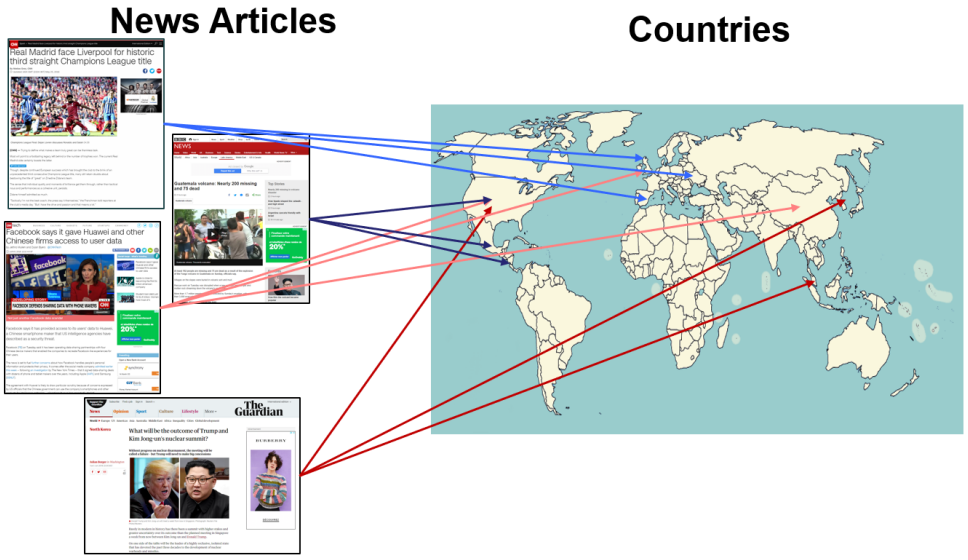


Figure 6.1: Interlinking the news stories to relevant countries

YAGO [Suchanek et al., 2007] or DBpedia [Auer et al., 2007], and methods for named entity disambiguation [Hoffart et al., 2011] we are able to exploit semantics of Web contents automatically and interpret them accordingly. To this end, we aim to build models, which are able to map the news articles to countries based on their relevance (cf. Figure 6.1).

In this chapter, we introduce ELEVATE-live, an extension of our ELEVATE framework [Govind and Spaniol, 2017], providing a Web-based user interface allowing its users an entity-level assessment and visualization in order to explore the interdependencies between Web news articles and geo-locations. To this end, our work makes the following contributions by:

- incorporating the ELEVATE framework and raising Web contents to the entity-level for semantic analytics;
- exploiting KBs in order to reveal non-trivial interdependencies between named entities contained and associated countries;
- providing a Web interface to study the "virality" of news articles with respect to countries concerned and vice versa.

6.1 Overview on ELEVATE-live

ELEVATE-live is a conceptual enhancement of the ELEVATE framework (cf. Chapter 4, [Govind and Spaniol, 2017]) allowing the assessment and visualization of Web contents by the example of online news articles. To this end, we "semantify" Web contents by harnessing location information associated with named entities and aggregating them for further analytics. The ultimate step is an analytics interface that allows exploring the "virality" w.r.t. the associated countries. Figure 6.2 highlights the five steps of data processing in ELEVATE-live, which will be explained subsequently.

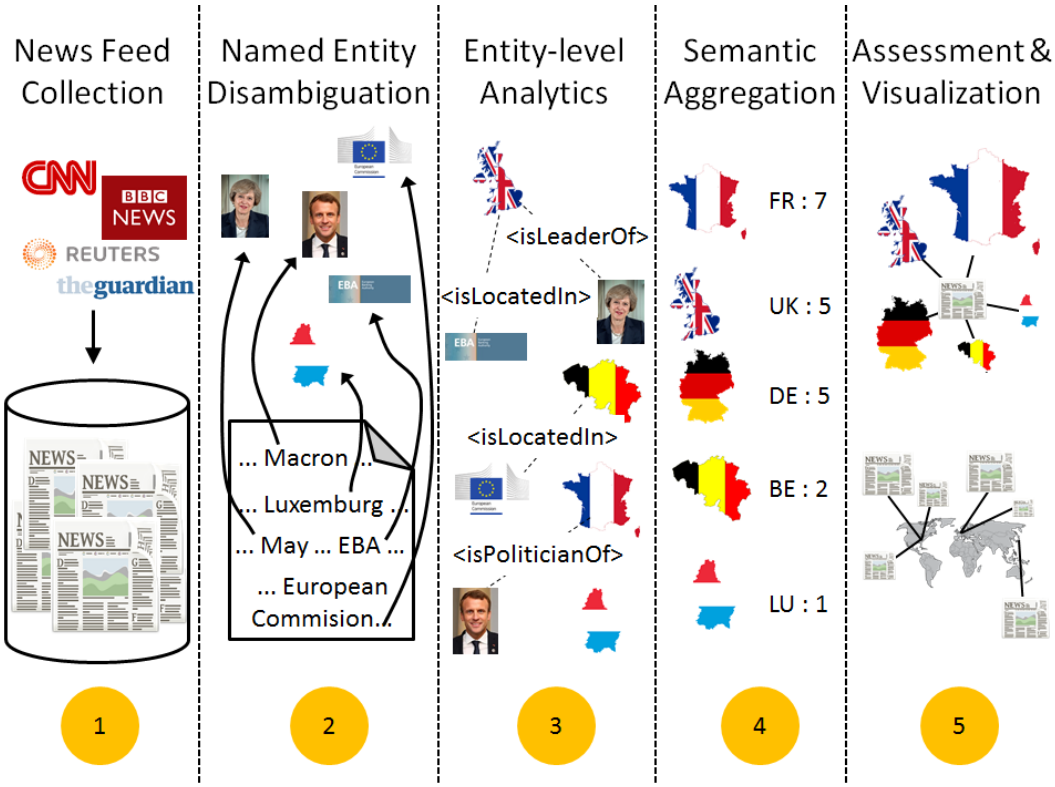


Figure 6.2: Conceptual approach of the ELEVATE-live pipeline illustrated by a `Brexit` related news article

6.1.1 News Feed Collection

In an initial step, we monitor the feeds of various online news agencies such as CNN, BBC, Reuters, etc. and fetch the latest news articles from these news sources. The fetched news articles are preprocessed to clean the HTML (Hypertext Markup Language) markup and the unwanted noise such as advertisements and other unrelated contents in the web page. We employ the standard boilerplate removal techniques to get the relevant text from web page of the news articles. Having the properly cleaned content, is important as the inclusion of irrelevant content or the exclusion of relevant content, can impact the system results. The impact can become severe when the original content is smaller in size.

6.1.2 Named Entity Extraction and Disambiguation

Subsequently, we employ AIDA [Hoffart et al., 2011] in order to reveal the named entities contained in a news article. By doing so, we raise each article to the entity-level. To this end, we obtain a list of canonical entities corresponding to each of the news articles. We also provide an option of using the Stanford named entity recognition system [Finkel et al., 2005], which can provide faster retrieval of results but have a trade-off with disambiguation accuracy of named entity mentions.

6.1.3 Entity-level Analytics

Next, the named entities contained in a news article are analyzed in order to gather location related information. To accomplish this, we utilize country- and organization-centric YAGO relations, such as `isLocatedIn`, `livesIn`, `worksAt`, etc. (refer to Chapter 4 for a complete list). As there are potentially many countries associated with a named entity (via different relations), the ELEVATE-live system pursue various strategies of knowledge base discovery:

1) Direct Mapping (DM): considering only the `country` mentions in the news article

2) Geographical location Mapping (GEO): considers only the `location` entities

3) Breadth-first-search (BFS): stopping when discovers an entity of type `country`

4) Depth-first-search (DFS): revealing all `countries` associated with a named entity

As aforementioned, ELEVATE framework provides several knowledge base exploration strategies to find the connections of the entities with the countries. There are two relatively simple strategies namely, direct mapping and geographical location mapping. We have two additional strategies which explores a wide variety of relations namely, the closest country discovery and the full exploration of countries. The first strategy is based on breadth first search where we search for the linked countries level by level and stops the exploration at the level where first country is discovered. The other strategy is based on the depth first search where we search exhaustively to discover all possible connections of an entity to different countries. The first strategy is less computationally expensive but the second one can uncover deep connections to the countries. Finally in result, we get a multiset containing relevant countries corresponding to each news article.

6.1.4 Semantic Aggregation

After that, we aggregate the geo-centric entity information derived from the previous step. Depending on the chosen exploitation strategy we obtain a set of associated `countries` associated with each article. Since, there are (usually) multiple relations associated with each named entity, this might lead to one (in the case of BFS, DM, GEO) or - potentially - many (in the case of DFS) associated countries per entity.

This is possible because there can be more than one paths to a country using separate entities in the news article. Here, we score each country by the number of times it is discovered while performing exploration for all the entities. This aggregated score reflects the importance of a country for the concerned news article, i.e. the high value signifies more relevance.

6.1.5 Countries Prediction

Now, the final task is to actually predict the countries where a news has the potential to get viral. In this final step, we provide a Web interface to assess and visualize the news articles. For this purpose, we rank all the candidate countries by their aggregated

semantic score as previously computed. We use a threshold based approach to filter out the candidates of low relevance. A user defined parameter θ is used to take the top θ countries from the ranked list. To this end, we utilize the extracted geo-information in order to rank and present the articles based on their relevance to specific countries or allow a country-based exploration of the most relevant articles.

6.2 Analytics Interface

ELEVATE-live facilitates the end user to assess and visualize the virality of news articles (cf. https://elevate.greyc.fr). In the following, we describe the two main use-cases of our system, i.e., individual news virality assessment, and exploration with respect to countries.

6.2.1 Assessing Virality from Semantically Enriched News

In our second use-case, the most recent news articles from these news sources are mapped on a zoomable timeline. Each news article is represented by a colored square. This news timeline allows the user to navigate through the news stories along the temporal dimension by summarizing the story in focus. The user can further explore the countries in which a story has the potential to become viral based on the entities contained. As before, the user may also explore the different exploration methods for semantic enrichment. The results in terms of countries "affected by the virus" are highlighted on an interactive world map, with a color coding from blue to red representing the degree of virality (cf. Figure 6.3). In addition, we provide interactive visualization of semantic connections between the named entities in the article with corresponding countries.

6.2.2 Assessing Viral News Stories by Country

In our first use-case, news contents can be searched by their relevance for a country based on the named entities contained in the article. Further options allow the user, e.g., to investigate the underlying models (DM, GEO, BFS, or DFS). In addition, temporal constraints can be defined in order to focus the query onto a certain time-interval. The document ordering is then done based on the aggregated score derived from the semantically enriched documents (cf. Figure 6.4).

6.3 Experimental Evaluation

In order to measure the performance of ELEVATE-live, we perform a basic evaluation. We randomly selected a sample of 50 news articles and manually assessed the countries associated with them (gold standard). We conducted experiments using several prediction models. As explained in Section 6.1, the ELEVATE models differ in the underlying KB exploration strategy, one is based on depth first search (DFS) and the other is based on breadth first search (BFS). In addition, we have three other models in ELEVATE-live: a direct mapping model (DM), a geolocation based model (GEO) and a baseline random

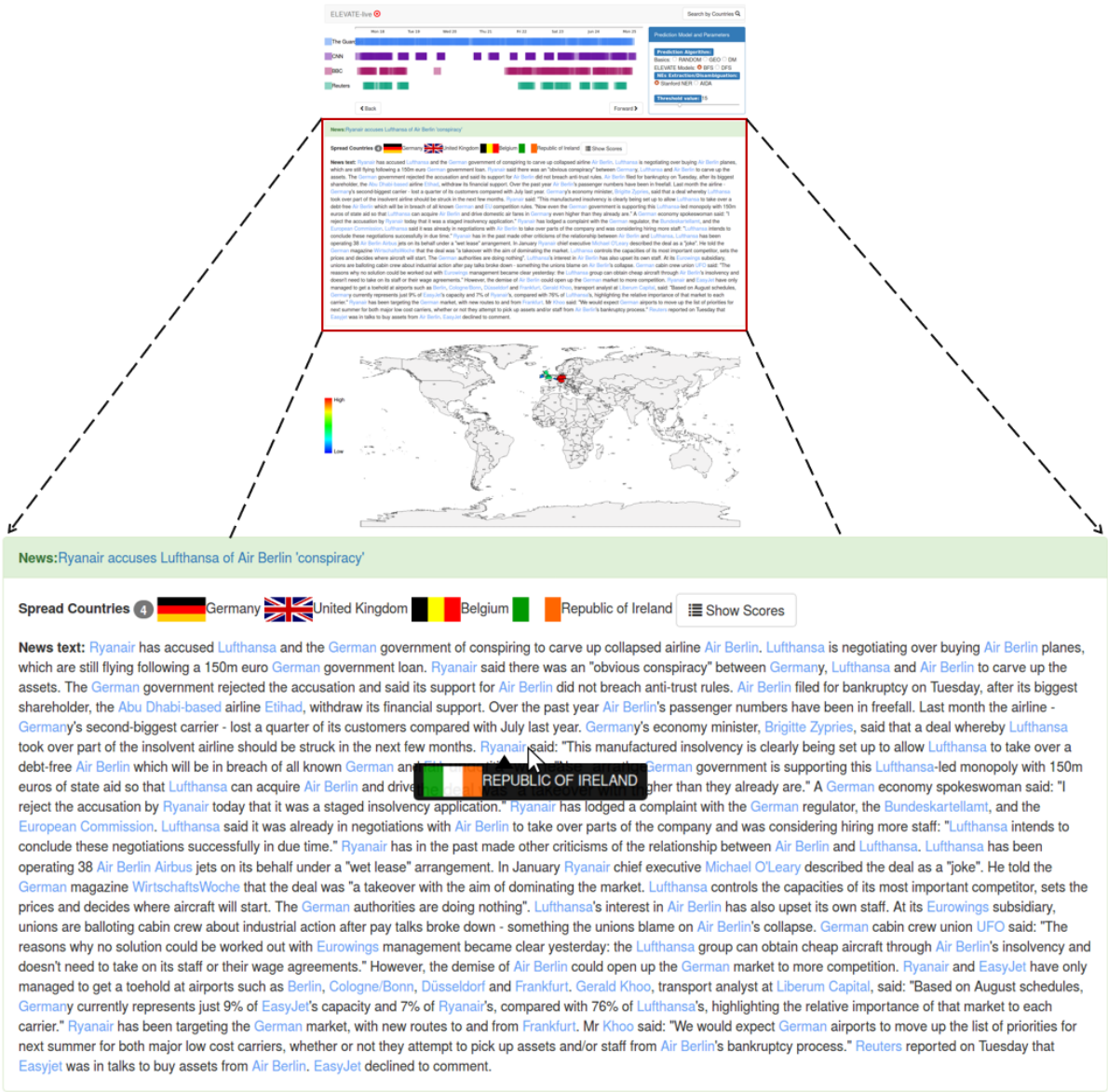


Figure 6.3: An illustration depicting the exploration of news article virality with the help of ELEVATE-live

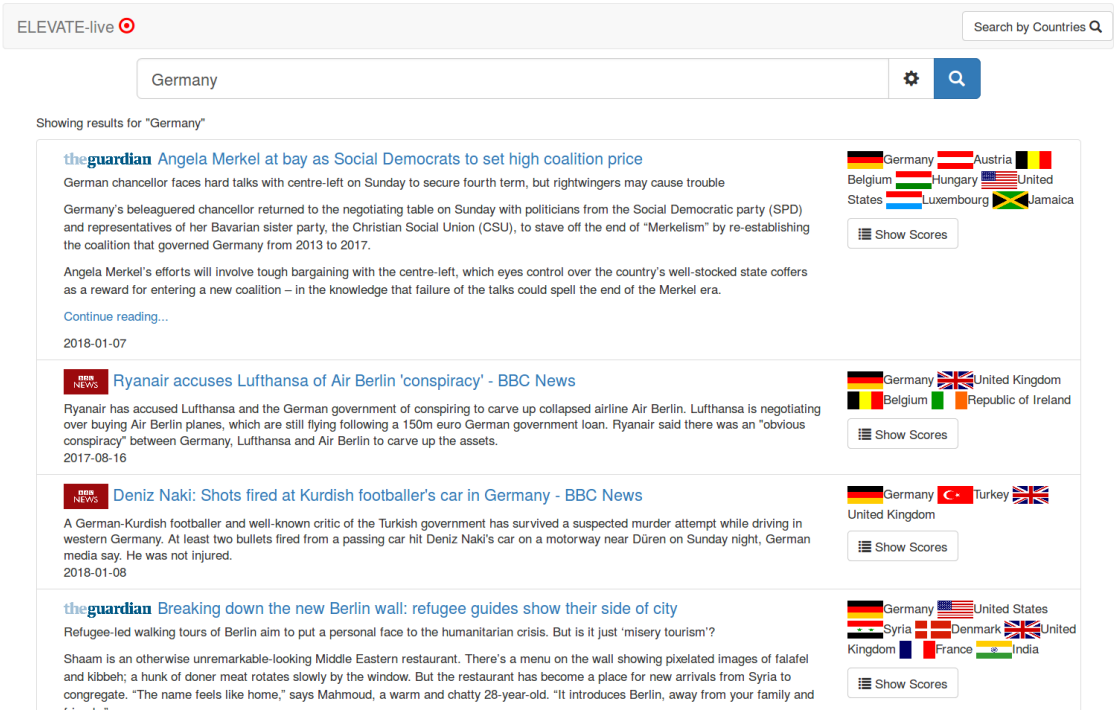


Figure 6.4: Country-specific viral/relevant news assessment

prediction model (Random). The DM consists in recognizing all the country names in a news article and predicting these countries. The GEO is a bit more sophisticated, it recognizes all the location entities (not only countries but cities, regions, etc.), derives the countries involved from these locations and predicts them. Finally, the random prediction model takes all the countries from the gold standard sample, computes the average number of countries associated with one article, and associates each article with this same number of countries (i.e. 3 for gold standard). The predicted countries for each article are selected randomly from all countries. Table 6.1 shows the results in terms of micro and macro precision, recall and F1-score.

As expected, the random prediction model has very low recall and precision. The two other approaches (DM and GEO) have a good precision[28]. Indeed, countries (or locations) mentioned in the news article are likely to be true positives, and since no other countries are predicted, there are not many false positives. On the opposite, for the ELEVATE models (BFS and DFS), more countries are usually predicted because it takes more information into account, meaning the number of false positives is higher, and this makes precision lower. For the same reason, the ELEVATE models (specifically BFS and DFS) outperform others in terms of recall. As more countries are discovered, so they are less false negatives.

When analyzing the country scores, we observed that the ELEVATE models have a bigger dispersion, e.g. going from 1 to 16 whereas for the baselines it is going from 1

[28] The DM makes very few predictions, so when no countries are predicted, precision is 0, which makes the macro-precision low.

to 6. This means that the ELEVATE models provide more accurate scores to discriminate among the slightly and the highly relevant countries.

Even if DM has a good precision, it misses many countries in its prediction, so the recall is very low. On the opposite, DFS considers more information, so its recall is high but precision goes low. We can conclude that the two best models are BFS and GEO, which have good F1-scores. BFS is the model to use when a good recall is needed, whereas GEO is the model to use when a good precision is needed.

	Micro			Macro		
Model	Prec.	Recall	F1	Prec.	Recall	F1
Random	0.0067	0.0063	0.0065	0.0067	0.004	0.005
DM	0.8256	0.4494	0.5820	0.4992	0.3493	0.4110
GEO	0.8095	0.6456	0.7183	0.7558	0.6998	0.7267
DFS	0.3069	0.7848	0.4413	0.3539	0.8321	0.4966
BFS	0.6122	0.7595	0.6780	0.6393	0.8130	0.7158

Table 6.1: Micro and macro-average scores for the ELEVATE-live prediction models

6.4 Findings on News Virality Analytics

In this chapter, we have presented a novel Web-based tool that exploits entity-level semantics driven by the ELEVATE framework in order to assess and visualize online news virality. The originality of our approach stems from analytics on the entity level. We observe that our system benefits from the information harvested via named entities associated, and is able to accurately reveal non-trivial interdependencies between news articles and countries. We performed a basic evaluation over a test set of news stories to examine ELEVATE-live usability and performance. We observe that it successfully captures the notion of virality/relevance from the geographical perspective. We enable end users to choose from a variety of options for news virality assessment models based on their individual priorities (such as fast retrieval, high recall or precision value). In addition, our analytics and visualization system comes with two modules. Here, one module is targeted to assess and visualize the virality with respect to individual news articles, and the other provides flexibility to address the problem from the other way, i.e., searching relevant news with respect to countries. A challenge that ELEVATE-live might face in some cases, is regarding the unavailability of emerging entities related information in knowledge base. This can be more relevant only in cases of news articles, which involve many newly emerging entities. However, we observe that these out of the scope scenarios are infrequent, and the system is effectively usable in majority of cases and has proven its practicality.

Chapter 7

Conclusion and Outlook

In this chapter, we conclude our findings from the research work done as part of this dissertation, and provide insights into the ongoing and future research.

As mentioned before, our society is observing an ever-growing virtual presence on the Web, which allows a variety of computational methods in order to study societal events. This thesis focused on the understanding of societal events on the Web, and investigated their different aspects such as spread, type, virality, etc. A noteworthy facet of the methods proposed here, is that they exploit extensively the semantic information derived via entity-level analytics. From our experiments we observe that the use of entity-level analytics furnishes a deeper understanding of Web documents, and a diverse set of tasks can consume this knowledge to perform better. We have proven with the help of ample experimentation that our formulated hypotheses (cf. Chapter 1) are successfully viable.

7.1 Findings on Entity-level Event Impact Analytics

In this section, we reflect upon our major findings on the work done as part of this thesis. In Chapter 4, we explore the problem of event diffusion prediction into foreign language communities, and have introduced a novel entity-based prediction framework ELEVATE for the same. We conducted a comprehensive study on events extracted from Wikipedia spanning over almost two decades. The accomplished experiments show very encouraging results. One of the salient contributions of our approach is the purely semantic analytics of Web contents. By doing so, we noticed a variety of benefits such as a reduction in the number of features, and a better document understanding over the bag-of-words model.

The ELEVATE framework relies solely on entity-level analytics without incorporating any additional link-based features, e.g., inter-language links from Wikipedia. This is particularly beneficial because it does not require some kind of "full-world knowledge" which is not explicitly available for Web contents per se. On the contrary, incorporating LOD via entities associated with the Web content is an efficient and scalable solution. We hypothesized that there should be intrinsic patterns in the spread originating from the homophily among the language communities (cf. Hypothesis 1 in Chapter 1). Our machine learning based model has been able to capture typical diffusion patterns by providing accurate spread predictions. We noticed the limitation of our model in a few use cases of predicting events spread, caused by the open and collaborative nature of the Web. In certain cases, passionate users/editors translate certain event pages to their native language or other languages of their expertise without any observable "connection" between event and language. For example, the existence of `2010 French pension reform strikes` page in Thai language. In such cases, there does not exist any detectable semantic connection to the target language. And thus, it becomes a challenge for our model to make a prediction. We observed that these cases are not very rampant in practice and thus, do not hamper the usability of our approach.

In the subsequent study, our experiments address the task of Web content classification with respect to a fine-grained hierarchy (cf. Chapter 5). We hypothesized that a document can be characterized by the named entities it contains (refer to Hypothesis 2 in Chapter 1). To this end, we proposed the "semantic fingerprinting" method that distills a Web content, and produces a concise representative vector by capturing the overall semantics of the document. We have shown the effectiveness of "semantic fingerprints" by performing experiments on a large dataset for the task of content alignment onto a fine-grained type hierarchy. The "semantic fingerprinting" method outperforms state-of-the-art competitors as it is able to successfully capture a semantic representation of Web contents. In this study, we observe that the type information related to the named entities contained in a Web content, carry concise knowledge about document core semantics. As a consequence, these information can be exploited to generate a semantic representation for Web contents. In addition, the "semantic fingerprinting" approach is scalable and can be adapted to any target application specific type hierarchy. It has been able to handle noise (which is common in case of Web contents) with the help of aggregation (averaging-out effect) over named entities. Moreover, a test example becomes more prone to noise when it is very small in size. It is worth pointing out that there exist several types in our type hierarchy which are inherently hard to distinguish, and thus, are more affected to misclassification by our model. For example, in case of types `football player` and `club`, it is hard to find a clear separation by only relying on named entities contained. However, even after the presence of such beyond scope factors, the semantic fingerprint approach provides very encouraging overall results.

Finally, in Chapter 6, we address the virality of online news articles. We hypothesized that the semantic analysis driven by ELEVATE framework can help in revealing various characteristics of online news virality (cf. Hypothesis 3 in Chapter 1). We have proven the viability of our proposed hypothesis by introducing ELEVATE-live, which has shown to be useful in assessing the relevance of online news articles. We observe that the ELEVATE framework can be effectively adapted to reveal the hidden semantic interde-

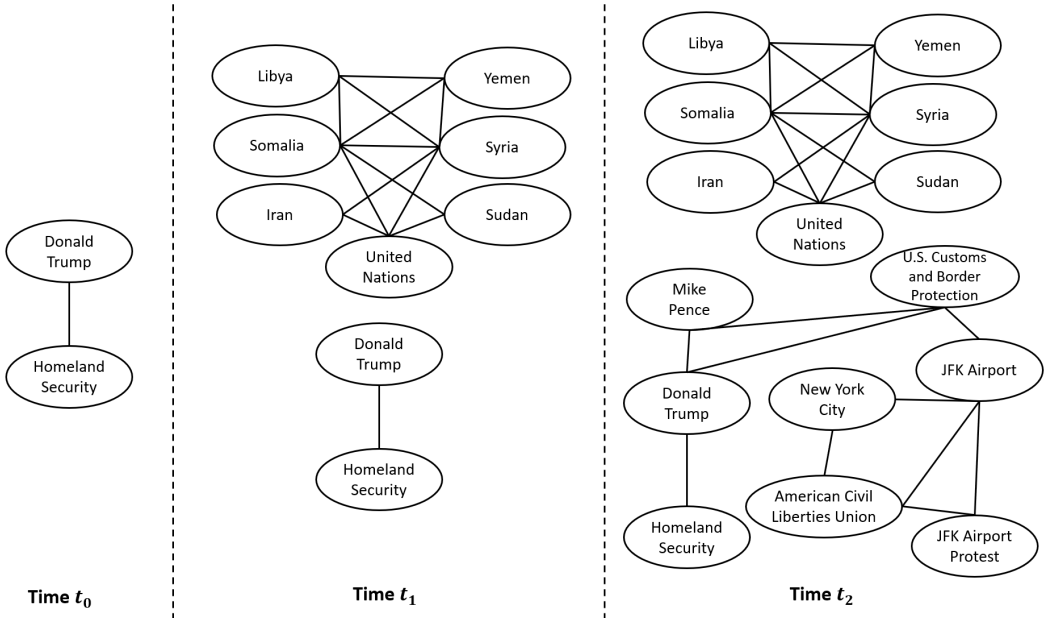
pendencies between the geographical regions and online news. Moreover, consistent with the findings from our event spread prediction and content classification work, entity-level analytics of the news articles has shown to unravel non-evident semantic relationships between news articles and "related" countries. For example, ELEVATE-live can reveal the relevant countries as well as their order of importance for the news article "*Ryanair accuses Lufthansa of Air Berlin conspiracy*"[29], i.e., Germany, United Kingdom, Belgium, and Republic of Ireland. It has been able to do so by exploiting named entities such as `Lufthansa`, `Ryanair`, `European Commission`, `Easyjet`, etc.

In addition to the discussed findings, we also observe some limitations, which are beyond the scope of work done in this dissertation. By the inherent nature of Web contents, there can exist uncontrolled noise such as markup code, advertisements, and other unrelated information. Although we perform standard preprocessing steps, this noise can still propagate, and have effects on the prediction performance of our models in some cases. In order to incorporate the entity-level features, our approaches involve a named entity recognition and disambiguation step. For certain studies disambiguation results were directly available via the markup (e.g. Wikipedia), while in the case of online news, an active disambiguation is required. Although NERD systems have seen many recent improvements in performance, but incorrect disambiguations can not be fully ruled out. Thus, the disambiguation error can propagate and affect the overall accuracy of our model. Similar to the NERD systems, automatically constructed KBs are not also completely error-free. In addition, they usually experience delays in adding emerging entities and keeping up-to-date information about the existing ones. As a result, the accuracy and coverage of information in KBs may affect the prediction capability of our model. Nevertheless, our approaches have proven great usability as the aforementioned limitations arise infrequently and are not the norm.

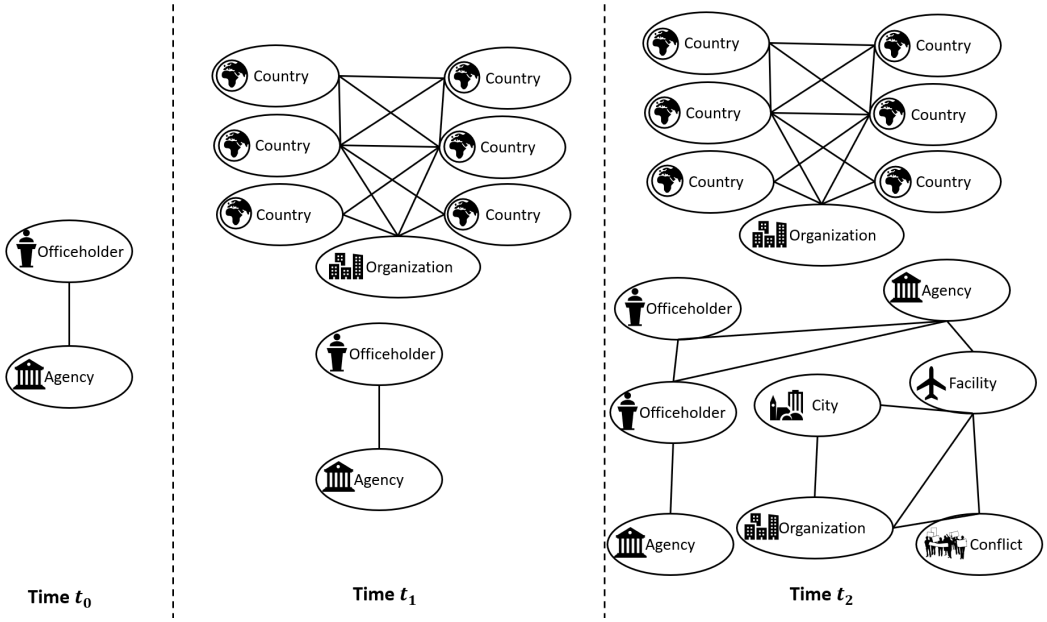
7.2 Ongoing Work - Patterns in Event Evolution

In order to predict various future characteristics (such as the type, spread, impact, and scale) of events early on, we are currently studying patterns in their evolution. We hypothesize that events of similar nature share the pattern in their evolution. To this end, we analyze the entity interaction/co-occurrence graph resulting from developments about an event along the temporal dimension. The graph is constructed with entities as nodes, and edges represented by the co-occurrence of two entities in the same sentence. Figure 7.1 depicts an illustration of entity co-occurrence graph at three different time points for the event `Muslim Travel Ban`. Figure 7.1a reflects to the entity co-occurrence in the temporal dimension, and Figure 7.1b shows the same graph at type-level. Initial experiments are currently being conducted on a diverse set of events belonging to around 50 major event categories extracted from Wikipedia. In these preliminary experiments, we observed that triads, cliques, or in general, dense subgraphs in the aforementioned graphs at different time points, appear to carry useful semantics about the evolution of event. We aim to explore the essence of an event by capturing these patterns over time. Also,

[29] https://www.bbc.com/news/technology-40949243



(a) Entity co-occurrence graphs depicting the evolution of an event over time



(b) Type representation of entity co-occurrence graphs depicting the evolution of an event over time

Figure 7.1: An illustration of the evolution of the event `Muslim Travel Ban`

we plan to incorporate recent techniques such as entity and graph embeddings [Lin et al., 2015, Grover and Leskovec, 2016] in matching the patterns of evolution among events.

7.3 Future Research Directions

The research work conducted in this thesis, inspires to pursue studies in diverse directions. In their respective chapters, we already suggested specific improvements on the approaches presented in our work. Here, we provide an overview of several of the major future directions resulting from our overall research.

7.3.1 Discovery and Explanation of Societal Perception

In the midst of so many daily societal activities, it is important to know "what can happen next?" as discussed previously, but it is of equal significance to know "why did something happen?". It is an inherently complicated task to automatically connect the dots to past happenings in the society, though it carries utmost value. In future, we aim at developing techniques that can provide the explanation and reasons behind a certain societal happening/perception. As discussed in Chapter 3, there have been several studies on targeting the prediction of future scenarios, but we intend to address the problem from the other way around. Given the proceedings of event happenings in the society, we plan to build a model that can generate the possible prerequisites that have led to the current scenario. A potential approach might investigate the usability of Granger causality in capturing such causal relationship between social happenings [Granger, 1969].

7.3.2 Exploration of Event Impact Aspects

In the current work, we pursue the problem of impact prediction of events from a foreign language community and/or geographical perspective. But there can be various other aspects to the impact of an event, such as its impact on economy, tourism, jobs, politics, international relations, etc., to name a few. Moreover, even a single event can have impact of varying degrees over multitude of sectors/domains. For example, in case of the event `Brexit`, there can be seen observable consequences on a variety of sectors in the affected geographical regions. The impact of `Brexit` potentially includes bilateral UK relations, immigration, economy, etc. To this end, we plan to explore the aspect based impact of events. In order to accomplish that, models have to be developed that can predict and quantify the impact of events over diverse societal domains. For this purpose, comprehensive experiments studying events with impact of global scale in a holistic manner will be required.

7.3.3 Disinformation Spread Detection

In recent times, our society is witnessing a rapidly growing dependence on the Web, which makes the Web an easy target for the wrongful or criminal deception for personal and/or financial gains. Disinformation on the Web has been flourishing because of various

reasons, prominent factors are online advertising revenue and tortious political influence. In recent context of major global events, there have been reports of undesirable influence caused by the spread of disinformation. Thus, the question of eminent importance is "how do we assess the credibility of contents on the Web?". As a natural next step to the research work done in this dissertation, entity-level analytics and inter-linkage with Linked Open Data might be studied in order to reveal and - in an ideal setting - help to explain/counter-argue disinformation. Therefore, techniques and an online interface should be developed in order to provide the credibility assessment of Web contents and information sources.

Appendix A

Countries to Major Language Mapping

Country	Language	Country	Language
Albania	Albanian	Sov. Mil. Order of Malta	Italian
Algeria	Arabic	Republic of Macedonia	Macedonian
Andorra	Catalan	Madagascar	French
Angola	Portuguese	Malawi	English
Antigua and Barbuda	English	Malaysia	Malaysian
Argentina	Spanish	Mali	French
Armenia	Armenian	Malta	English
Australia	English	Marshall Islands	English
Austria	German	Mauritania	Arabic
Azerbaijan	Azerbaijani	Mauritius	English
The Bahamas	English	Mexico	Spanish
Bahrain	Arabic	Fed. States of Micronesia	English
Bangladesh	Bengali	Moldova	Romanian
Barbados	English	Monaco	French
Belarus	Belarusian	Montenegro	Albanian
Belgium	Dutch	Morocco	Arabic
Belize	English	Mozambique	Portuguese
Benin	French	Myanmar	Burmese
Bolivia	Spanish	Nagorno-Karabakh	Armenian
Bosnia and Herzegovina	Bosnian	Namibia	English
Botswana	English	Nauru	English
Brazil	Portuguese	Nepal	Nepali
Brunei	Malay	Netherlands	Dutch
Bulgaria	Bulgarian	New Zealand	English
Burkina Faso	French	Nicaragua	Spanish
Burundi	English	Niger	French
Cameroon	English	Nigeria	English

Canada	English	Northern Cyprus	Turkish
Cabo Verde	Portuguese	Norway	Bokmål
Central African Republic	French	Oman	Arabic
Chad	Arabic	Pakistan	Urdu
Chile	Spanish	Palau	English
China	Chinese	State of Palestine	Arabic
Colombia	Spanish	Panama	Spanish
Comoros	Arabic	Papua New Guinea	English
Dem. Rep. of the Congo	French	Paraguay	Spanish
Republic of the Congo	French	Peru	Spanish
Costa Rica	Spanish	Philippines	English
Croatia	Croatian	Poland	Polish
Cuba	Spanish	Portugal	Portuguese
Cyprus	Greek	Qatar	Arabic
Czechia	Czech	Romania	Romanian
Denmark	Danish	Russia	Russian
Djibouti	Arabic	Rwanda	English
Dominica	English	Sahrawi Arab Dem. Rep.	Arabic
Dominican Republic	Spanish	Saint Kitts and Nevis	English
East Timor	Portuguese	Saint Lucia	English
Ecuador	Spanish	Sa. Vi. and the Grenadines	English
Egypt	Arabic	Samoa	English
El Salvador	Spanish	San Marino	Italian
Equatorial Guinea	Spanish	São Tomé and Príncipe	Portuguese
Eritrea	Arabic	Saudi Arabia	Arabic
Estonia	Estonian	Scotland	Scots
Ethiopia	English	Senegal	French
Fiji	English	Serbia	Serbian
Finland	Finnish	Seychelles	English
France	French	Sierra Leone	English
Gabon	French	Singapore	English
The Gambia	English	Slovakia	Slovak
Georgia (country)	Georgian	Slovenia	Hungarian
Germany	German	Solomon Islands	English
Ghana	English	Somalia	Arabic
Greece	Greek	Somaliland	Arabic
Grenada	English	South Africa	Afrikaans
Guatemala	Spanish	South Ossetia	Russian
Guinea	French	South Sudan	English
Guinea-Bissau	Portuguese	Spain	Spanish
Guyana	English	Sri Lanka	Tamil
Haiti	French	Sudan	Arabic

Honduras	Spanish	Suriname	Dutch
Hungary	Hungarian	Swaziland	English
Iceland	Icelandic	Sweden	Swedish
India	Hindi	Switzerland	German
Indonesia	Indonesian	Syria	Arabic
Iran	Persian	Taiwan	Chinese
Iraq	Arabic	Tajikistan	Tajik
Republic of Ireland	English	Tanzania	Swahili
Israel	Hebrew	Thailand	Thai
Italy	Italian	Togo	French
Ivory Coast	French	Transnistria	Russian
Jamaica	English	Tonga	English
Japan	Japanese	Trinidad and Tobago	English
Jordan	Arabic	Tunisia	Arabic
Kazakhstan	Kazakh	Turkey	Turkish
Kenya	English	Turkmenistan	Russian
Kiribati	English	Tuvalu	English
North Korea	Korean	Uganda	English
South Korea	Korean	Ukraine	Ukrainian
Kosovo	Albanian	United Arab Emirates	Arabic
Kuwait	Arabic	United Kingdom	English
Kyrgyzstan	Kirghiz	United States	English
Latvia	Latvian	Uruguay	Spanish
Lebanon	Arabic	Uzbekistan	Uzbek
Lesotho	English	Vanuatu	English
Liberia	English	Vatican City	Latin
Libya	Arabic	Venezuela	Spanish
Liechtenstein	German	Vietnam	Vietnamese
Lithuania	Lithuanian	Yemen	Arabic
Luxembourg	French	Zambia	English
Macau	Cantonese	Zimbabwe	English

Appendix B

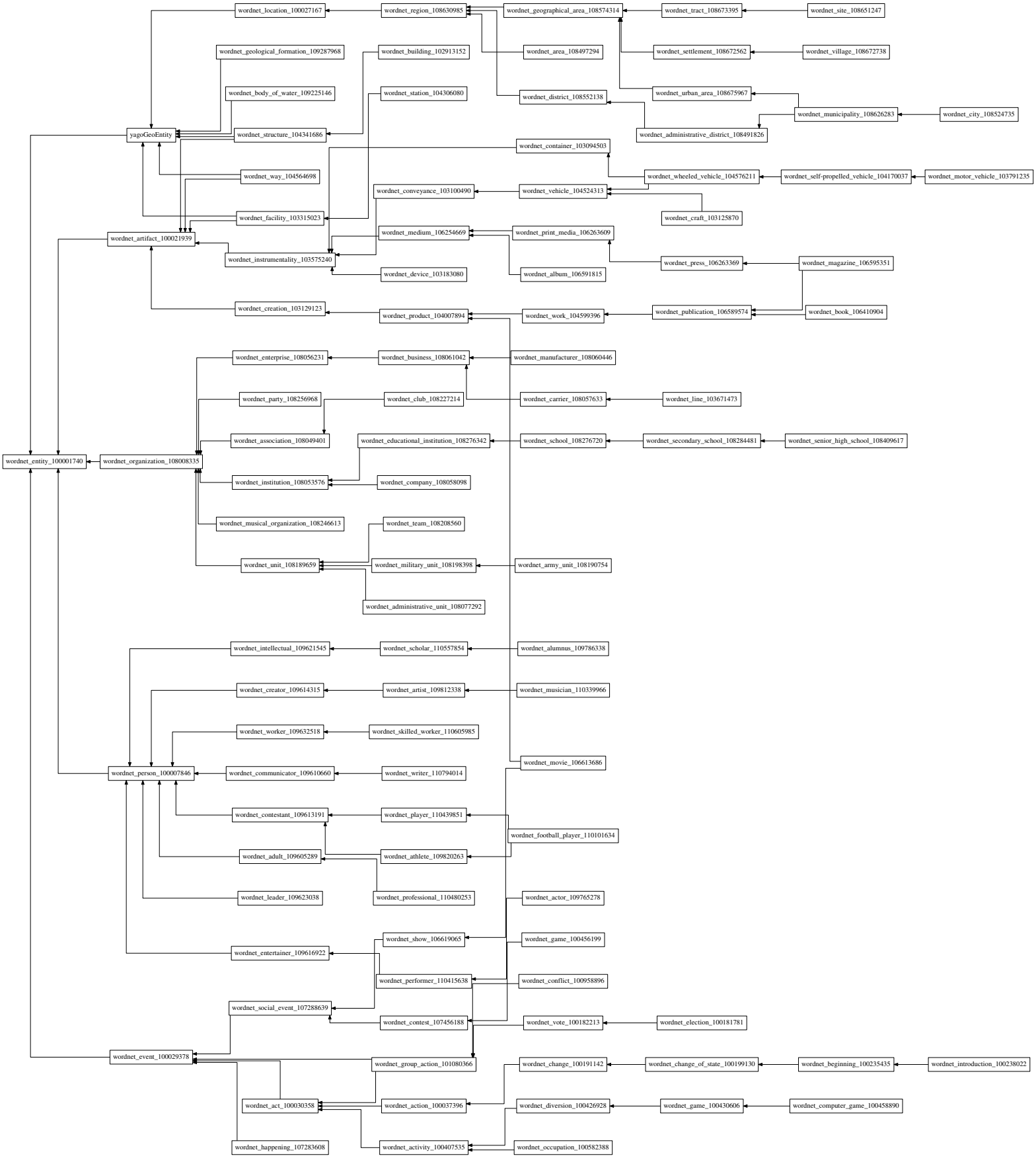
Abbreviations

Abbreviation	Full Form	Page
AA	Adamic Adar coefficient	22
ACE	Automatic Contect Extraction	27
ACM	Association of Computating Machinary	8
AIDA	Accurate Online Disambiguation of Named Entities	14
API	Application Programming Interface	57
ARIMA	Auto-regressive Integrated Moving Average	31
BFS	Breadth First Search	51
CN	Common Neighbors	22
CNN	Convolutional Neural Network	27
CRF	Conditional Random Field	37
DAG	Directed Acyclic Graph	69
DFS	Depth First Search	51
DM	Direct Mapping	80
DNN	Deep Neural Network	39
ELEVATE	Entity-LEVel AnalyTics for Event diffusion prediction framework	6
ELEVATE-live	Entity-LEVel AnalyTics for Event diffusion prediction framework - live	8
ERE	Entities Relations Events	27
FIGER	FIne-Grained Entity Recognition	37
FN	False Negative	23
FP	False Positive	23
GDELT	Global Database of Events, Language, and Tone	31
GEO	Geographical location mappping	80
GMM	Gaussian Mixture Model	26
GRU	Gated Recurrent Unit	27
HTML	Hypertext Markup Language	79
HYENA	Hierarchical tYpe classification for Entity NAmes	37
ICF	Inverse Candidate Frequency	15
ICT	Information and Communication Technology	1

ICWE	Internation Conference on Web Engineering	8
IR	Information Retrieval	35
JC	Jaccard's Coefficient	23
KB	Knowledge Base	5
k-NN	k-Nearest Neighbors	37
LASSO	Least Absolute Shrinkage and Selection Operator	31
LDA	Latent Dirichlet Allocation	31
LOD	Linked Open Data	5
LSTM	Long-Short Term Memory	27
MAP	Maximum A Posteriori	17
MaxEnt	Maximum Entropy	29
MSTParser	Maximum Spanning Tree Parser	29
NB	Naive Bayes	17
NCI	News Cohesiveness Index	32
NED	Named Entity Disambiguation	14
NERC	Named Entity Recognition and Classification	14
NERD	Named Entity Recognition and Disambiguation	14
NLP	Natural Language Processing	5
NLTK	Natural Language Toolkit	14
NYT	New York Times	31
PA	Preferential Attachment	23
PC	Prediction Count	59
PoS	Part of Speech	37
RBF	Radial Basis Function	37
RDF	Resource Data Framework	12
RDFS	Resource Data Framework Schema	13
RF	Random Forest	18
RNN	Recurrent Neural Network	29
STICS	Searching with Strings, Things, and Cats	34
SVM	Support Vector Machine	19
tf-icf	Term Frequency-Inverse Candidate Frequency	15
tf-idf	Term Frequency-Inverse Document Frequency	17
TN	True Negative	23
TP	True Positive	23
URI	Uniform Resource Identifier	43
URL	Uniform Resource Locator	27
W3C	World Wide Web Consortium	12
YAGO	Yet Another Great Ontology	5

Appendix C

Type Hierarchy


wordnet_location_100027167
wordnet_region_108630985
wordnet_geographical_area_108574314
wordnet_tract_108673395
wordnet_site_108651247
wordnet_geological_formation_109287968
wordnet_building_102913152
wordnet_area_108497294
wordnet_settlement_108672562
wordnet_village_108672738
wordnet_body_of_water_109225146
wordnet_station_104306080
wordnet_district_108552138
wordnet_urban_area_108675967
wordnet_municipality_108626283
wordnet_city_108524735
yagoGeoEntity
wordnet_structure_104341686
wordnet_administrative_district_108491826
wordnet_container_103094503
wordnet_way_104564698
wordnet_wheeled_vehicle_104576211
wordnet_self-propelled_vehicle_104170037
wordnet_motor_vehicle_103791235
wordnet_conveyance_103100490
wordnet_vehicle_104524313
wordnet_facility_103315023
wordnet_craft_103125870
wordnet_artifact_100021939
wordnet_medium_106254669
wordnet_print_media_106263609
wordnet_instrumentality_103575240
wordnet_press_106263369
wordnet_magazine_106595351
wordnet_device_103183080
wordnet_album_106591815
wordnet_creation_103129123
wordnet_product_104007894
wordnet_work_104599396
wordnet_publication_106589574
wordnet_book_106410904
wordnet_enterprise_108056231
wordnet_business_108061042
wordnet_manufacturer_108060446
wordnet_party_108256968
wordnet_club_108227214
wordnet_carrier_108057633
wordnet_line_103671473
wordnet_association_108049401
wordnet_educational_institution_108276342
wordnet_school_108276720
wordnet_secondary_school_108284481
wordnet_senior_high_school_108409617
wordnet_entity_100001740
wordnet_organization_108008335
wordnet_institution_108053576
wordnet_company_108058098
wordnet_musical_organization_108246613
wordnet_team_108208560
wordnet_unit_108189659
wordnet_military_unit_108198398
wordnet_army_unit_108190754
wordnet_administrative_unit_108077292
wordnet_intellectual_109621545
wordnet_scholar_110557854
wordnet_alumnus_109786338
wordnet_creator_109614315
wordnet_artist_109812338
wordnet_musician_110339966
wordnet_worker_109632518
wordnet_skilled_worker_110605985
wordnet_movie_106613686
wordnet_person_100007846
wordnet_communicator_109610660
wordnet_writer_110794014
wordnet_contestant_109613191
wordnet_player_110439851
wordnet_football_player_110101634
wordnet_adult_109605289
wordnet_athlete_109820263
wordnet_leader_109623038
wordnet_professional_110480253
wordnet_actor_109765278
wordnet_show_106619065
wordnet_game_100456199
wordnet_entertainer_109616922
wordnet_performer_110415638
wordnet_conflict_100958896
wordnet_social_event_107288639
wordnet_contest_107456188
wordnet_vote_100182213
wordnet_election_100181781
wordnet_event_100029378
wordnet_group_action_101080366
wordnet_change_100191142
wordnet_change_of_state_100199130
wordnet_beginning_100235435
wordnet_introduction_100238022
wordnet_act_100030358
wordnet_action_100037396
wordnet_diversion_100426928
wordnet_game_100430606
wordnet_computer_game_100458890
wordnet_happening_107283608
wordnet_activity_100407535
wordnet_occupation_100582388

List of Figures

List of Tables

Bibliography

[Adamic and Adar, 2003] Adamic, L. A. and Adar, E. (2003). Friends and neighbors on the web. *Social Networks*, 25(3):211 – 230.

[Ahn et al., 2011] Ahn, B. G., Van Durme, B., and Callison-Burch, C. (2011). Wikitopics: What is popular on wikipedia and why. In *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, WASDGML '11, pages 33–40, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Alec et al., 2016] Alec, C., Reynaud-Delaître, C., and Safar, B. (2016). An Ontology-Driven Approach for Semantic Annotation of Documents with Specific Concepts. In *The Semantic Web. Latest Advances and New Domains. 13th ESWC 2016*, pages 609–624, Heraklion, Greece. Springer.

[Allahyari et al., 2014] Allahyari, M., Kochut, K. J., and Janik, M. (2014). Ontology-based text classification into dynamically defined topics. In *2014 IEEE International Conference on Semantic Computing*, pages 273–278.

[Amsterdamer et al., 2013] Amsterdamer, Y., Grossman, Y., Milo, T., and Senellart, P. (2013). Crowd mining. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pages 241–252, New York, NY, USA. ACM.

[Angel et al., 2012] Angel, A., Sarkas, N., Koudas, N., and Srivastava, D. (2012). Dense subgraph maintenance under streaming edge weight updates for real-time story identification. *Proc. VLDB Endow.*, 5(6):574–585.

[Atefeh and Khreich, 2013] Atefeh, F. and Khreich, W. (2013). A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164.

[Auer et al., 2007] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. G. (2007). Dbpedia: A nucleus for a web of open data. In *ISWC/ASWC*, pages 722–735.

[Barbieri et al., 2014] Barbieri, N., Bonchi, F., and Manco, G. (2014). Who to follow and why: Link prediction with explanations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1266–1275, New York, NY, USA. ACM.

[Bastos et al., 2015] Bastos, M. T., Mercea, D., and Charpentier, A. (2015). Tents, tweets, and events: The interplay between ongoing protests and social media. *Journal of Communication*, 65(2):320–350.

[Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific american*, 284(5):34–43.

[Bird et al., 2009] Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

[Bizer et al., 2009] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). Dbpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154 – 165. The Web of Data.

[Bizer et al., 2011] Bizer, C., Heath, T., and Berners-Lee, T. (2011). Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227. IGI Global.

[Bollacker et al., 2008] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250, New York, NY, USA. ACM.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[Cadena et al., 2015] Cadena, J., Korkmaz, G., Kuhlman, C. J., Marathe, A., Ramakrishnan, N., and Vullikanti, A. (2015). Forecasting social unrest using activity cascades. *PLOS ONE*, 10(6):1–27.

[Cano et al., 2013] Cano, A. E., Varga, A., Rowe, M., Ciravegna, F., and He, Y. (2013). Harnessing linked knowledge sources for topic classification in social media. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, HT '13, pages 41–50, New York, NY, USA. ACM.

[Chakraborty et al., 2016] Chakraborty, S., Venkataraman, A., Jagabathula, S., and Subramanian, L. (2016). Predicting socio-economic indicators using news events. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1455–1464, New York, NY, USA. ACM.

[Chaney et al., 2016] Chaney, A., Wallach, H., Connelly, M., and Blei, D. (2016). Detecting and characterizing events. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1142–1152.

[Chang and Lin, 2011] Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[Chen and Roy, 2009] Chen, L. and Roy, A. (2009). Event detection from flickr data through wavelet-based spatial analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 523–532. ACM.

[Cheng et al., 2014] Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., and Leskovec, J. (2014). Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 925–936, New York, NY, USA. ACM.

[Ciampaglia et al., 2015] Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., and Flammini, A. (2015). Computational fact checking from knowledge networks. *PLOS ONE*, 10(6):1–13.

[Clauset Aaron et al., 2008] Clauset Aaron, Moore Cristopher, and Newman M. E. J. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101. 10.1038/nature06830.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

[Dalton et al., 2014] Dalton, J., Dietz, L., and Allan, J. (2014). Entity query feature expansion using knowledge base links. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 365–374, New York, NY, USA. ACM.

[de Loupy et al., 1998] de Loupy, C., Bellot, P., El-Bèze, M., and Marteau, P.-F. (1998). Query expansion and classification of retrieved documents. In *TREC*, pages 382–389.

[Dong et al., 2015] Dong, Y., Zhang, J., Tang, J., Chawla, N. V., and Wang, B. (2015). Coupledlp: Link prediction in coupled networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 199–208, New York, NY, USA. ACM.

[Dos Santos et al., 2016] Dos Santos, R. F., Boedihardjo, A., Shah, S., Chen, F., Lu, C.-T., and Ramakrishnan, N. (2016). The big data of violent events: algorithms for association analysis using spatio-temporal storytelling. *GeoInformatica*, 20(4):879–921.

[Elberrichi et al., 2008] Elberrichi, Z., Rahmoun, A., and Bentaallah, M. A. (2008). Using WordNet for Text Categorization. *Int. Arab J. Inf. Technol.*, 5:16–24.

[Fang and Ben-Miled, 2017] Fang, A. and Ben-Miled, Z. (2017). Does bad news spread faster? In *2017 International Conference on Computing, Networking and Communications (ICNC)*, pages 793–797.

[Feng et al., 2018] Feng, X., Qin, B., and Liu, T. (2018). A language-independent neural network for event detection. *Science China Information Sciences*, 61(9):092106.

[Ferragina and Scaiella, 2010] Ferragina, P. and Scaiella, U. (2010). TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*, pages 1625–1628.

[Fetahu et al., 2015] Fetahu, B., Anand, A., and Anand, A. (2015). How much is wikipedia lagging behind news? In *Proceedings of the ACM Web Science Conference*, page 28. ACM.

[Finkel et al., 2005] Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics.

[Firth, 1957] Firth, J. (1957). *A Synopsis of Linguistic Theory, 1930-1955*.

[Fleischman and Hovy, 2002] Fleischman, M. and Hovy, E. (2002). Fine grained classification of named entities. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Freire et al., 2016] Freire, A., Manca, M., Saez-Trumper, D., Laniado, D., Bordino, I., Gullo, F., and Kaltenbrunner, A. (2016). Graph-based breaking news detection on wikipedia. *Wiki Workshop, ICWSM 2016*, 6:1.

[Gao et al., 2010] Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., and Zhao, B. Y. (2010). Detecting and characterizing social spam campaigns. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, IMC '10, pages 35–47, New York, NY, USA. ACM.

[García, 2015] García, F. F. (2015). Analyzing and visualizing news spread based on images in social media networks. *MS Thesis, University of Amsterdam*.

[Govind and Spaniol, 2017] Govind and Spaniol, M. (2017). ELEVATE: A Framework for Entity-level Event Diffusion Prediction into Foreign Language Communities. In *Proceedings of the 9th International ACM Web Science Conference (WebSci '17)*, pages 111–120.

[Govind et al., 2018a] Govind, Alec, C., and Spaniol, M. (2018a). ELEVATE-Live: Assessment and Visualization of Online News Virality via Entity-Level Analytics. In *Proceedings of the 18th International Conference on Web Engineering, ICWE 2018, Cáceres, Spain, June 5-8, 2018*, pages 482–486.

[Govind et al., 2018b] Govind, Alec, C., and Spaniol, M. (2018b). Semantic Fingerprinting: A Novel Method for Entity-Level Content Classification. In *Proceedings of the 18th International Conference on Web Engineering, ICWE 2018, Cáceres, Spain, June 5-8, 2018*, pages 279–287.

[Govind, 2018] Govind (2018). Entity-level event impact analytics. *WSTNET Web Science Summer School, Hannover, Germany, July 30 - Aug 4, 2018*.

[Granger, 1969] Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.

[Grover and Leskovec, 2016] Grover, A. and Leskovec, J. (2016). Node2vec: Scalable feature learning for networks. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 855–864, New York, NY, USA. ACM.

[Hansen et al., 2011] Hansen, L. K., Arvidsson, A., Nielsen, F. A., Colleoni, E., and Etter, M. (2011). *Good Friends, Bad News - Affect and Virality in Twitter*, pages 34–43. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Hashimoto et al., 2014] Hashimoto, C., Torisawa, K., Kloetzer, J., Sano, M., Varga, I., Oh, J.-H., and Kidawara, Y. (2014). Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 987–997. Association for Computational Linguistics.

[Ho, 1998] Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(8):832–844.

[Hoffart et al., 2011] Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Thater, S., and Weikum, G. (2011). Robust disambiguation of named entities in text. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 782–792.

[Hoffart et al., 2013] Hoffart, J., Suchanek, F. M., Berberich, K., and Weikum, G. (2013). YAGO2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, 194:28–61.

[Hoffart et al., 2014] Hoffart, J., Milchevski, D., and Weikum, G. (2014). Stics: Searching with strings, things, and cats. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 1247–1248, New York, NY, USA. ACM.

[Hong et al., 2011] Hong, Y., Zhang, J., Ma, B., Yao, J., Zhou, G., and Zhu, Q. (2011). Using cross-entity inference to improve event extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1127–1136, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Hotho et al., 2003] Hotho, A., Staab, S., and Stumme, G. (2003). Ontologies improve text document clustering. In *Proceedings of the Third IEEE International Conference on Data Mining*, ICDM '03, pages 541–, Washington, DC, USA. IEEE Computer Society.

[Huet et al., 2013] Huet, T., Biega, J., and Suchanek, F. M. (2013). Mining history with le monde. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 49–54, New York, NY, USA. ACM.

[Ifrim et al., 2014] Ifrim, G., Shi, B., and Brigadir, I. (2014). Event detection in twitter using aggressive filtering and hierarchical tweet clustering. In *Second Workshop on Social News on the Web (SNOW), Seoul, Korea, 8 April 2014*. ACM.

[Jaccard, 1901] Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579.

[Jatowt et al., 2015] Jatowt, A., Antoine, E., Kawai, Y., and Akiyama, T. (2015). Mapping temporal horizons: Analysis of collective future and past related attention in twitter. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 484–494, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

[Jenders et al., 2013] Jenders, M., Kasneci, G., and Naumann, F. (2013). Analyzing and predicting viral tweets. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion, pages 657–664, New York, NY, USA. ACM.

[Joachims, 1998] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In Nédellec, C. and Rouveirol, C., editors, *Machine Learning: ECML-98*, pages 137–142, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Johnson and Zhang, 2014] Johnson, R. and Zhang, T. (2014). Effective use of word order for text categorization with convolutional neural networks. *CoRR*, abs/1412.1058.

[Joulin et al., 2016] Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

[Kallus, 2014] Kallus, N. (2014). Predicting crowd behavior with big public data. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 625–630, New York, NY, USA. ACM.

[Kalman, 1960] Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45.

[Keneshloo et al., 2016] Keneshloo, Y., Wang, S., Han, E.-H. S., and Ramakrishnan, N. (2016). Predicting the popularity of news articles. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 441–449.

[Kim and Leskovec, 2011] Kim, M. and Leskovec, J. (2011). *The Network Completion Problem: Inferring Missing Nodes and Edges in Networks*, pages 47–58. SIAM.

[Kim, 2014] Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

[Lai et al., 2015] Lai, S., Xu, L., Liu, K., and Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *AAAI*, volume 333, pages 2267–2273.

[Li et al., 2016] Li, H., Ellis, J. G., Ji, H., and Chang, S.-F. (2016). Event specific multimodal pattern mining for knowledge base construction. In *Proceedings of the 2016 ACM on Multimedia Conference*, MM '16, pages 821–830, New York, NY, USA. ACM.

[Li et al., 2013] Li, Q., Ji, H., and Huang, L. (2013). Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 73–82.

[Li et al., 2005] Li, Z., Wang, B., Li, M., and Ma, W.-Y. (2005). A probabilistic model for retrospective news event detection. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 106–113, New York, NY, USA. ACM.

[Liao and Grishman, 2010] Liao, S. and Grishman, R. (2010). Using document level cross-event inference to improve event extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 789–797, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Liben-Nowell and Kleinberg, 2007] Liben-Nowell, D. and Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031.

[Lilleberg et al., 2015] Lilleberg, J., Zhu, Y., and Zhang, Y. (2015). Support vector machines and word2vec for text classification with semantic features. In *2015 IEEE 14th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC)*, pages 136–140.

[Lin et al., 2015] Lin, Y., Liu, Z., Sun, M., Liu, Y., and Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion. In *AAAI*, volume 15, pages 2181–2187.

[Ling and Weld, 2012] Ling, X. and Weld, D. S. (2012). Fine-grained entity recognition. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, pages 94–100. AAAI Press.

[Lü and Zhou, 2011] Lü, L. and Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150 – 1170.

[Lütkebohle, 2008] Lütkebohle, I. (2008). Distributed RDF Dataset Statistics. `http://sansa-stack.net/distlodstats/`. [Online; accessed 17-August-2018].

[Manning et al., 2008] Manning, C. D., Raghavan, P., Schütze, H., et al. (2008). *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.

[Matsubara et al., 2015] Matsubara, Y., Sakurai, Y., and Faloutsos, C. (2015). The web as a jungle: Non-linear dynamical systems for co-evolving online activities. In *Proceedings of the 24th International Conference on World Wide Web*, pages 721–731. International World Wide Web Conferences Steering Committee.

[McClosky et al., 2011] McClosky, D., Surdeanu, M., and Manning, C. D. (2011). Event extraction as dependency parsing. In *Proceedings of the 49th Annual Meeting of the*

Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11, pages 1626–1635, Stroudsburg, PA, USA. Association for Computational Linguistics.

[McIver and Brownstein, 2014] McIver, D. J. and Brownstein, J. S. (2014). Wikipedia usage estimates prevalence of influenza-like illness in the united states in near real-time. *PLoS Comput Biol*, 10(4):e1003581.

[Mendes et al., 2011] Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, pages 1–8, New York, NY, USA. ACM.

[Michalski et al., 2013] Michalski, R. S., Carbonell, J. G., and Mitchell, T. M. (2013). *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.

[Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

[Miller, 1995] Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

[Milne and Witten, 2008] Milne, D. and Witten, I. H. (2008). Learning to link with wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 509–518, New York, NY, USA. ACM.

[Moreno et al., 2017] Moreno, J. G., Besançon, R., Beaumont, R., D'hondt, E., Ligozat, A.-L., Rosset, S., Tannier, X., and Grau, B. (2017). Combining word and entity embeddings for entity linking. In *European Semantic Web Conference*, pages 337–352. Springer.

[Muthiah et al., 2015] Muthiah, S., Huang, B., Arredondo, J., Mares, D., Getoor, L., Katz, G., and Ramakrishnan, N. (2015). Planned protest modeling in news and social media. In *AAAI*, pages 3920–3927.

[Nadeau and Sekine, 2007] Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticæ Investigationes*, 30(1):3–26.

[Newman, 2001] Newman, M. E. J. (2001). Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 64:025102.

[Nguyen et al., 2016] Nguyen, T. H., Cho, K., and Grishman, R. (2016). Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309.

[Osborne et al., 2012] Osborne, M., Petrovic, S., McCreadie, R., Macdonald, C., and Ounis, I. (2012). Bieber no more: First story detection using twitter and wikipedia. In *SIGIR 2012 Workshop on Time-aware Information Access*. ACM.

[Petrović et al., 2010] Petrović, S., Osborne, M., and Lavrenko, V. (2010). Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 181–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Peñalver-Martinez et al., 2014] Peñalver-Martinez, I., Garcia-Sanchez, F., Valencia-Garcia, R., Ángel Rodríguez-García, M., Moreno, V., Fraga, A., and Sánchez-Cervantes, J. L. (2014). Feature-based opinion mining through ontologies. *Expert Systems with Applications*, 41(13):5995 – 6008.

[Piškorec et al., 2014] Piškorec, M., Antulov-Fantulin, N., Novak, P. K., Mozetič, I., Grčar, M., Vodenska, I., and Šmuc, T. (2014). Cohesiveness in financial news and its relation to market volatility. *Scientific reports*, 4:5038.

[Quinlan, 1993] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Radinsky and Horvitz, 2013] Radinsky, K. and Horvitz, E. (2013). Mining the web to predict future events. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 255–264, New York, NY, USA. ACM.

[Rahman and Ng, 2010] Rahman, A. and Ng, V. (2010). Inducing fine-grained semantic classes via hierarchical and collective classification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 931–939, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Rifkin and Klautau, 2004] Rifkin, R. and Klautau, A. (2004). In defense of one-vs-all classification. *J. Mach. Learn. Res.*, 5:101–141.

[Rozenshtein et al., 2014] Rozenshtein, P., Anagnostopoulos, A., Gionis, A., and Tatti, N. (2014). Event detection in activity networks. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1176–1185, New York, NY, USA. ACM.

[Sakaki et al., 2010] Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA. ACM.

[Sebastiani, 2002] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47.

[Song et al., 2011] Song, Y., Wang, H., Wang, Z., Li, H., and Chen, W. (2011). Short text conceptualization using a probabilistic knowledgebase. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, pages 2330–2336. AAAI Press.

[Strube and Ponzetto, 2006] Strube, M. and Ponzetto, S. P. (2006). Wikirelate! computing semantic relatedness using wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, pages 1419–1424. AAAI Press.

[Suchanek et al., 2007] Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). YAGO: A core of semantic knowledge - unifying WordNet and Wikipedia. In *16th International World Wide Web Conference (WWW 2007)*, pages 697–706. ACM.

[Suchanek and Preda, 2014] Suchanek, F. M. and Preda, N. (2014). Semantic culturomics. *Proc. VLDB Endow.*, 7(12):1215–1218.

[Tang et al., 2009] Tang, L., Rajan, S., and Narayanan, V. K. (2009). Large scale multi-label classification via metalabeler. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 211–220, New York, NY, USA. ACM.

[Tsoumakas and Katakis, 2006] Tsoumakas, G. and Katakis, I. (2006). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3).

[Usbeck et al., 2014] Usbeck, R., Ngonga Ngomo, A.-C., Röder, M., Gerber, D., Coelho, S. A., Auer, S., and Both, A. (2014). *AGDISTIS - Graph-Based Disambiguation of Named Entities Using Linked Data*, pages 457–471. Springer International Publishing, Cham.

[W3C et al., 2014] W3C et al. (2014). Rdf 1.1 concepts and abstract syntax.

[Wang et al., 2016] Wang, C., Song, Y., Li, H., Zhang, M., and Han, J. (2016). Text classification with heterogeneous information network kernels. In *AAAI*, pages 2130–2136.

[Wang et al., 2015] Wang, P., Xu, B., Wu, Y., and Zhou, X. (2015). Link prediction in social networks: the state-of-the-art. *Science China Information Sciences*, 58(1):1–38.

[Weikum et al., 2011] Weikum, G., Ntarmos, N., Spaniol, M., Triantafillou, P., Benczúr, A. A., Kirkpatrick, S., Rigaux, P., and Williamson, M. (2011). Longitudinal analytics on web archive data: It's about time! In *CIDR*, pages 199–202.

[Weng et al., 2013] Weng, L., Menczer, F., and Ahn, Y.-Y. (2013). Virality prediction and community structure in social networks. *Scientific reports*, 3:2522.

[Whiting et al., 2014] Whiting, S., Jose, J., and Alonso, O. (2014). Wikipedia as a time machine. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 857–862, New York, NY, USA. ACM.

[Xu et al., 2018] Xu, J., Mei, T., Cai, R., Li, H., and Rui, Y. (2018). Automatic generation of social event storyboard from image click-through data. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(1):242–253.

[Yan et al., 2015] Yan, Y., Yang, Y., Meng, D., Liu, G., Tong, W., Hauptmann, A. G., and Sebe, N. (2015). Event oriented dictionary learning for complex event detection. *IEEE Transactions on Image Processing*, 24(6):1867–1878.

[Yang and Counts, 2010] Yang, J. and Counts, S. (2010). Predicting the speed, scale, and range of information diffusion in twitter. *Icwsm*, 10(2010):355–358.

[Yang et al., 1999] Yang, Y., Carbonell, J. G., Brown, R. D., Pierce, T., Archibald, B. T., and Liu, X. (1999). Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems and their Applications*, 14(4):32–43.

[Yang et al., 2016] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

[Yosef et al., 2011] Yosef, M. A., Hoffart, J., Bordino, I., Spaniol, M., and Weikum, G. (2011). AIDA: An Online Tool for Accurate Disambiguation of Named Entities in Text and Tables. In *Proc. of the* 37^{th} *Intl. Conference on Very Large Databases (VLDB 2011), August 29 - September 3, Seattle, WA, USA*, pages 1450–1453.

[Yosef et al., 2012] Yosef, M. A., Bauer, S., Hoffart, J., Spaniol, M., and Weikum, G. (2012). Hyena: Hierarchical type classification for entity names. In *Proceedings of COLING 2012: Posters*, pages 1361–1370. The COLING 2012 Organizing Committee.

[Yosef et al., 2013] Yosef, M. A., Bauer, S., Hoffart, J., Spaniol, M., and Weikum, G. (2013). HYENA-live: Fine-Grained Online Entity Type Classification from Natural-language Text. In *51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, Proceedings of the Conference System Demonstrations, 4-9 August 2013, Sofia, Bulgaria*, pages 133–138. The Association for Computer Linguistics.

[Zhang et al., 2017] Zhang, T., Whitehead, S., Zhang, H., Li, H., Ellis, J., Huang, L., Liu, W., Ji, H., and Chang, S.-F. (2017). Improving event extraction via multimodal integration. In *Proceedings of the 2017 ACM on Multimedia Conference*, MM '17, pages 270–278, New York, NY, USA. ACM.

Entity-level Event Impact Analytics

Abstract: Our society has been rapidly growing its presence on the Web, as a consequence we are digitizing a large collection of our daily happenings. In this scenario, the Web receives virtual occurrences of various events corresponding to their real world occurrences from all around the world. Scale of these events can vary from locally relevant ones up to those that receive global attention. News and social media of current times provide all essential means to reach almost a global diffusion. This big data of complex societal events provide a platform to many research opportunities for analyzing and gaining insights into the state of our society.

In this thesis, we investigate a variety of social event impact analytics tasks. Specifically, we address three facets in the context of events and the Web, namely, diffusion of events in foreign languages communities, automated classification of Web contents, and news virality assessment and visualization. We hypothesize that the named entities associated with an event or a Web content carry valuable semantic information, which can be exploited to build accurate prediction models. We have shown with the help of multiple studies that raising Web contents to the entity-level captures their core essence, and thus, provides a variety of benefits in achieving better performance in diverse tasks. We report novel findings over disparate tasks in an attempt to fulfill our overall goal on societal event impact analytics.

Keywords: Societal Events Analysis, Entity-level Web Analytics, Multilingual Web Data, Semantically-enriched Web Content Classification, Web Semantics

Analyse de l'Impact des Événements au Niveau des Entités

Résumé: Notre société est de plus en plus présente sur le Web. En conséquence, une grande partie des événements quotidiens a vocation à être numérisée. Dans ce cadre, le Web contient des descriptions de divers événements du monde réel et provenant du monde entier. L'ampleur de ces événements peut varier, allant de ceux pertinents uniquement localement à ceux qui retiennent l'attention du monde entier. La presse et les médias sociaux permettent d'atteindre une diffusion presque mondiale. L'ensemble de toutes ces données décrivant des événements sociétaux potentiellement complexes ouvre la porte à de nombreuses possibilités de recherche pour analyser et mieux comprendre l'état de notre société.

Dans cette thèse, nous étudions diverses tâches d'analyse de l'impact des événements sociétaux. Plus précisément, nous abordons trois facettes dans le contexte des événements et du Web, à savoir la diffusion d'événements dans des communautés de langues étrangères, la classification automatisée des contenus Web et l'évaluation et la visualisation de la viralité de l'actualité. Nous émettons l'hypothèse que les entités nommées associées à un événement ou à un contenu Web contiennent des informations sémantiques précieuses, qui peuvent être exploitées pour créer des modèles de prédiction précis. À l'aide de nombreuses études, nous avons montré que l'élévation du contenu Web au niveau des entités saisissait leur essence essentielle et offrait ainsi une variété d'avantages pour obtenir de meilleures performances dans diverses tâches. Nous exposons de nouvelles découvertes sur des tâches disparates afin de réaliser notre objectif global en matière d'analyse de l'impact des événements sociétaux.

Mots-clés: Analyse d'événements sociétaux, Analyse du Web au niveau des entités, Données Web multilingues, Classification de contenus Web sémantiquement enrichis, Sémantique du Web